THE ROLE OF EARLY RECURRENCE IN IMPROVING VISUAL REPRESENTATIONS

XUN SHI

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Graduate Program in
Computer Science and Engineering
York University
Toronto, Ontario
January 2016

Abstract

This dissertation proposes a computational model of early vision with recurrence, termed as early recurrence. The idea is motivated from the research of the primate vision. Specifically, the proposed model relies on the following four observations. 1) The primate visual system includes two main visual pathways: the dorsal pathway and the ventral pathway; 2) The two pathways respond to different visual features; 3) The neurons of the dorsal pathway conduct visual information faster than that of the neurons of the ventral pathway; 4) There are lower-level feedback connections from the dorsal pathway to the ventral pathway. As such, the primate visual system may implement a recurrent mechanism to improve visual representations of the ventral pathway.

Our work starts from a comprehensive review of the literature, based on which a conceptualization of early recurrence is proposed. Early recurrence manifests itself as a form of surround suppression. We propose that early recurrence is capable of refining the ventral processing using results of the dorsal processing.

Our work further defines a set of computational components to formalize early recurrence. Although we do not intend to model the true nature of biology, to verify that the proposed computation is biologically consistent, we have applied the model to simulate a neurophysiological experiment of a bar-and-checkerboard and a psychological experiment involving a moving contour illusion. Simulation results indicated that the proposed computation behaviourally reproduces the original observations.

The ultimate goal of this work is to investigate whether the proposal is capable of improving computer vision applications. To do this, we have applied the model to a variety of applications, including visual saliency and contour detection. Based on comparisons against the state-of-the-art, we conclude that the proposed model of early recurrence sheds light on a generally applicable yet lightweight approach to boost real-life application performance.

Acknowledgement

I would like to express my sincerest thanks to my supervisor, Professor John Tsotsos, for his continued support, encouragement and guidance with great patience from day to day. Along this long journey, he not only supervised me to finish this dissertation, but also guided me towards an independent researcher, an idea presenter, and an explorer, without which I would not reach my goal. To him and his wife, Pat, for making an excellent role model for me and my family.

To my parents and my in-laws for supporting me as always, and for selflessly backing me up across the Pacific Ocean (from China to Canada).

To my lovely wife and my soul mate, Lele, for taking care of me, for allowing me playing silly and childish, and for igniting my passion to excel.

To my little angel Amber and my little hero Jasper, for keeping me staying young, and for reminding me constantly that I need to grow together with them.

To my friends for enriching my life, and for informing me about the outside world.

Last but not least, to my committee members for providing valuable suggestions and helpful feedbacks that are important to my work.

Contents

List of Tables

List of Figures

List of Acronyms

AUC:	Area under curve
BM:	Benchmark
CS:	Centre-surround
DoG:	Difference-of-Gaussians
ER:	Early recurrence/recurrent
IT:	Inferior temporal
LGN:	Lateral geniculate nucleus
MGC:	Midget ganglion cell
MST:	Medial superior temporal
MT:	Middle temporal
PGC:	Parasol ganglion cell
PR:	Precision-recall
RF:	Receptive field
RGC:	Retina ganglion cell
ROC:	Receiver operating characteristic
WTA:	Winner-take-all

Chapter 1. Introduction

The dominant paradigm in computer vision today is that of a feed-forward process that uses hierarchical layers of representations, determined mostly by statistical learning methods, that extract feature vectors that are then passed through a classifier for ultimate decisions on image content. The process has met with substantial success as a variety of high profile results have demonstrated (Babaud et al. 1986, Biederman 1987, Cheng et al. 1998, Lowe 1999, Riesenhuber and Poggio 1999, Angelucci and Sainsbury 2006, Fei-Fei et al. 2006, Serre et al. 2007, Bay et al. 2008). In essence, this is a classic pattern recognition strategy as has been popular for decades (Fidler and Leonardis 2007). Successes are difficult to criticize, but when such successes are coupled with claims of faithfulness to biological vision, then there is room for debate. Although it is true that hierarchies and layered representations are part of biological visual processing, it is important to add that much more about biological vision is simply ignored in this dominant approach. One of those characteristics is feedback or recurrence. In a variety of other engineering domains, feedback is a well-understood topic, effectively used for a long time (Distefano III et al. 1967). However, in biological vision, even though it is acknowledged that feedback is present, its role remains an open question.

A strong message from this dissertation is that recurrent processes in computer vision are important, and that one of the roles they play is to provide spatial and temporal context to improve lower-level and early representations. In this sense, the term “lower-level” and “early” are interchangeably used. They refer to the very first few levels of visual areas in the primate visual system. In what follows, we will specify these areas in detail. At this point, readers can refer to Figure 3-1 to get a brief idea of areas referred as early visual areas. In addition, visual areas beyond these areas are higher-level or late visual areas in this work.

Quoting from (Felzenszwalb et al. 2010):

“In classical models of object recognition, first, basic features (e.g., edges and lines) are analyzed by independent filters that mimic the receptive field profiles of V1 neurons. In a feed-forward fashion, the outputs of these filters are fed to filters at the next processing stage, pooling information across several filters from the previous level, and so forth at subsequent processing stages. Lower-level processing determines higher-level processing. Information lost on lower stages is irretrievably lost. Models of this type have proven to be very successful in many fields of vision, but have failed to explain object recognition in general. Here, we present experiments that, first, show that, similar to demonstrations from the Gestaltists; figural aspects determine lower-level processing (as much as the other way around). Second, performance on a single element depends on all the other elements in the visual scene. Small changes in the overall configuration can lead to large changes in performance. Third, grouping of elements is the key. Only if we know how elements group across the entire visual field, can we determine performance on individual elements, i.e., challenging the classical stereotypical filtering approach, which is the very heart of most vision models.”

Such experimental support for our position is not new, but the above is among the most recent.

The primate visual system is a large neural network consisting of a massive number of feed-forward and feedback connections (Felleman and Van Essen 1991, Lienhart and Maydt 2002). Between the primary visual cortex (V1) and the lateral geniculate nucleus (LGN), feedback influences feed-forward processing in a pathway-specific fashion (i.e., the ventral pathway and the dorsal pathway) to sharpen the receptive fields of LGN neurons, and to enhance the transmission of signals through the LGN (Field 1987, Fogel and Sagi 1989, Georgeson et al. 2007). Within these visual areas, feedback modulatory effects have been widely observed. In one study (Angelucci and Bullier 2003), by inactivating high-order visual areas the authors noticed a major decrease in the strength of surrounding neurons in low-order visual areas. This hints at a mechanism where feedback plays a major role in centre-surround interactions. In (Jonas and Buzsaki 2007, Herzog and Clarke 2014), feedback circuits from the visual area MT (V5) to the visual area V1 and the thalamus have been shown to improve motion perception. In (Hupé et al. 1998, Jones et al. 2002), the authors studied feedback influence on figure-segregation. They showed that feedback processing amplifies activity of low-order neurons, particularly with low-visible stimuli. Feedback effects have been observed even without stimuli (Koenderink 1984, Keller et al. 2012).

1.1 Motivations

The first major motivation to our work is that that early visual areas do not simply act to transform feed-forward signals. In fact, they also integrate top-down and lateral information to refine visual representations. The underlying visual hierarchy is thus not a feed-forward only cascade, and neural activity is not simply data-driven, as proposed by (Hubel and Wiesel

1959, Hubel and Wiesel 1962).

As pointed out by a recent review (Kravitz et al. 2013), the ventral pathway is a recurrent occipitotemporal network containing neural representations to associate stimulus with response, to process emotionally salient stimuli, to support the assignment of stimulus valence, to support long-term memory, to support object-reward association, and to support object working memory. Neurons in the ventral pathway communicate via bidirectional connections. More specifically, the ventral pathway connects the early visual cortex to higher cortical structures to form object representations. The ventral pathway receives most feed-forward information of visual stimuli from the parvocellular layer of the LGN. The main visual processing areas include areas V1, V2, V4 and the inferior temporal (IT) cortex. Visual information enters the ventral pathway via area V1 and reaches the rest of the visual areas in sequence. Along the pathway, size of neuron's receptive field (region of the sensory space) gradually increases, with visual representation becoming complicated.

One of the functional roles of ventral processing is to generate view-invariant representations. Here we borrow the term "invariant" from computer vision, where it characterizes a statistic of an image that is stable to a well-defined set of image transformations, such as deformation and scale. A number of neuroscience studies have shown that neurons in the ventral pathway have view-invariant characteristics (Lueschow et al. 1994, Rolls 2000, Quiroga et al. 2005) for object and face recognition (Gobbini and Haxby 2007). Note that this invariant characteristic is in contrast to the dynamic relationship among objects within the scene, which is one of the roles of the dorsal processing.

Obviously, the ventral processing includes a set of mechanisms to extract and refine visual information to get view-invariance. Classical theories of the ventral pathway give rise to the idea that it is a serialized hierarchy, with each sequential stage having more complicated selectivity and invariance than its lower stage. This sequential view excludes feedback or lateral processing. However, recent studies suggest that the ventral pathway is in fact a recurrent network. In addition, the network has connections with the dorsal hierarchy at multiple stages. The interaction between the dorsal and ventral processing depends on a number of factors. To this point, we need to introduce the second motivation for the current work.

The second major motivating element for this thesis is Bullier's fast-brain hypothesis (Bullier 2001). The literature has not addressed this hypothesis in any detail with respect to the representations and algorithms that might realize it, neither within computer vision nor within computational neuroscience. Bullier argued that feedback connections are the best candidates for rapid long-distance communication. In his hypothesis, the LGN is the key area. It is a processing hub in the thalamus, which receives major sensory input from the retina, and relays most of the information to visual area V1. The LGN has three types of cells, the magnocellular cells, the parvocellular cells, and the koniocellular cells. Each type has distinct neurophysiological properties. Importantly, information conducted via the magnocellular cells reach visual cortices much earlier than that via the parvocellular cells. Bullier suggested that results from this first-pass computation are then sent back by feedback connections to areas V1 and V2, which act as “active black-boards” for the rest of the visual cortical areas: information retro-injected from the parietal cortex is used to guide further processing of parvocellular and koniocellular information in the inferior temporal cortex.

Bullier and his colleagues concluded that latency of visual processing at different visual cortices does not conform to a nice feed-forward only hierarchical pattern in the Hubel and

Wiesel sense (Nowak et al. 1997, Hupé et al. 1998, Hupé et al. 2001). Areas in the dorsal pathway, such as areas MT and MST, respond to input much faster than areas in the ventral pathway. Further, they noticed that there exist multitude cross-pathway recurrent connections from higher-level dorsal regions to lower-level ventral regions. Based on these observations, Bullier proposed that if timing difference permits, the visual system might utilize these recurrent connections to send results from the dorsal processing to modulate ventral processing.

Functional studies suggested that the fast-brain mechanism plays an active role in refining visual processing in the ventral pathway. A figure-ground discrimination study (Hupé et al. 1998) showed that feedbacks from the dorsal pathway increase the difference between neural responses to a bar moving on a stationary background and neural responses to the same bar moving together with the background. An fMRI study (Seghier et al. 2000a) using moving Kanisza illusory rectangles found strong and reproducible signal-boosting in areas V1 and V2, which are likely caused by recurrence from area MT/V5. In (Beck and Neumann 2010), the authors noted that MT-V1 feedback strengthens boundary perception in the chopstick illusion experiment. They further suggested that MT-V1 modulation facilitates localized motion estimation, which is not possible by pure feed-forward processing. Although evidence suggests that impacts of the fast-brain mechanism widely exist in the visual system, attempts to model it have not had much progress since Bullier’s early efforts. The significance of it to the big picture of biological vision and computer vision is not well understood.

The third and final motivation for this thesis is the scale-space theory (Witkin 1983). Specifically, one could draw a conceptual connection between the fast-brain hypothesis and the scale-space analysis. The core idea of scale-space analysis is to examine signals from a multi-scale perspective. In image processing, the basic operation is to convolve an image with a group of Gaussian kernels with different variances (scales). As the scale increases, the convolved image becomes coarser. This allows gradually smoothing out high-frequency image components (i.e., image noise and sharp edges). Stacking the convolved outputs along scales constructs a scale-space representation. It has a desired property of causality: once an edge disappears at one scale, it will not show in any coarser scales. As such, tracing the location of an edge coarse-to-fine becomes available.

Witkin used one-dimensional signals as an example and proposed an algorithm to compute "stable variances" across the scale space (Witkin 1983). He argued that edge response at stable variances has a marked correspondence with perceptually salient object contours. However, when variance increases, the scale-space representation will lose spatial accuracy, that edge response will shift from its actual location. To solve this loss-of-accuracy problem, anisotropic diffusion has been proposed to preserve spatial information during constructing the scale-space representation (Gregoriou et al. 2009).

Figure 1-1 illustrates a one-dimension example. Figure A shows the smoothed output signals as a result of applying Gaussian kernel progressively over the original signal. Clearly shown in figure B, signals are able to catch the intuitive notion of fine-scale information, or causality.

During the era of 1990, Lindeberg and his colleagues applied scale-space analysis to solve computer vision problems (Lindeberg 1991, Lindeberg 1993). In (Lindeberg 1994, Lindeberg 1998), the author addressed the problem of feature detection with automatic scale selection. By detecting local extrema over scales of differential expressions, scale selection

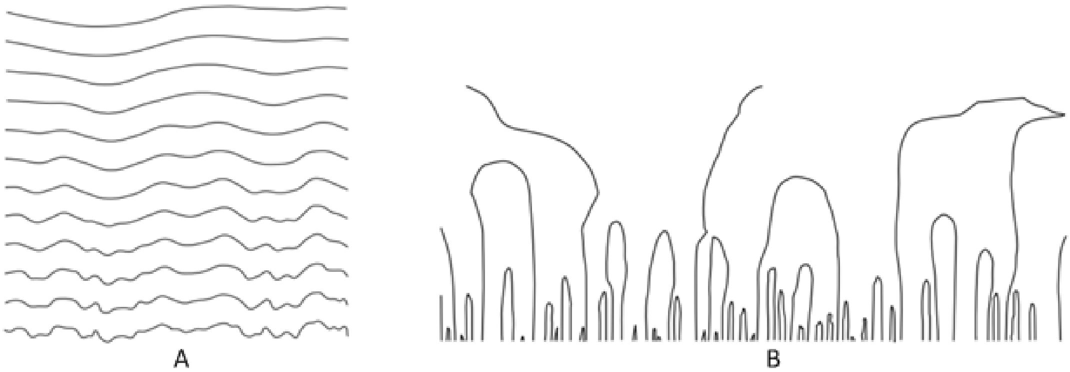


Figure 1-1 (A) A scale-space representation of one-parameter family of derived signals. The fine-scale information is progressively suppressed. (B) Gaussian smoothing does not create new zero-crossings, thus the trajectories of zero-crossings in scale-space are never closed from below. Image sourced from (Lindeberg 1996).

complements traditional scale-space theory by providing explicit mechanisms for predicting interesting scales. Figure 1-2 illustrates two examples of difficult edge detection that requires auto scale selection. In the bear example, the hairs are high-frequency pixel variations. It challenges the algorithm to pick the silhouette of the bear and ignore the hairs. In the sunflower example, it requires scale selection to choose coarser scales to process near-by sunflowers, and to choose finer scales to process far-away sunflowers.

In the literature, the discussion of the relationship between the scale-space analysis and the primate visual system is limited to an abstract level. It has been hypothesized that the primate visual system is capable of computing scale-space smoothing (Lindeberg 1996). Figure 1-3 depicts a hierarchy of receptive fields at which scale-space processing occurs. Overlaid circles represent receptive fields at each stage of the visual hierarchy. As scale increases, there is a significant increase of size of receptive fields: a neuron at a coarser scale sees larger spatial region.

There is one aspect of our use of scale-space that differs with the classical views and those expressed in the above figures. The above description is a representational scheme where the only parameter that changes as scale increases is kernel size; that is, the nature of what each kernel computes does not change (for example, each kernel detects edges of the same kind but at different scales). In our model, not only does the scale change along the visual hierarchy, but also the nature of the kernel: at an early layer edges may be computed, while at a higher layer object translation direction might be computed. At different scales, our model extracts different visual characteristics. This makes the direct application of previous scale-space methods inappropriate.

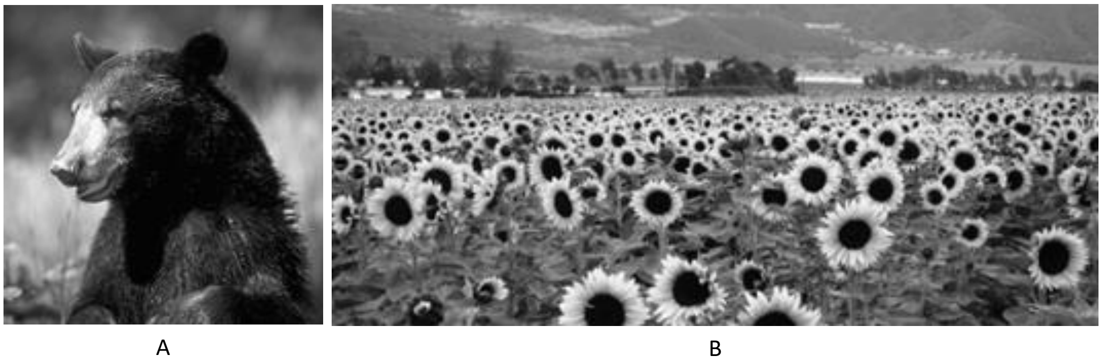


Figure 1-2 Examples of challenging edge detection. (A) To extract the silhouette of the bear body, hairs must be discarded (as image noise). (B) Nearby and far-away sunflowers require edge detection to be scale adaptive.

The presentational abstractions of the proposed hierarchy are further examined in Figure 1-4. The figure illustrates a 4-stage hierarchy from the finest scale to the coarsest scale. Coloured circles indicate different natures of kernel at each abstraction level. Although these kernels extract different types of visual features, their relative sizes across scales still satisfy the scale-space constraints.

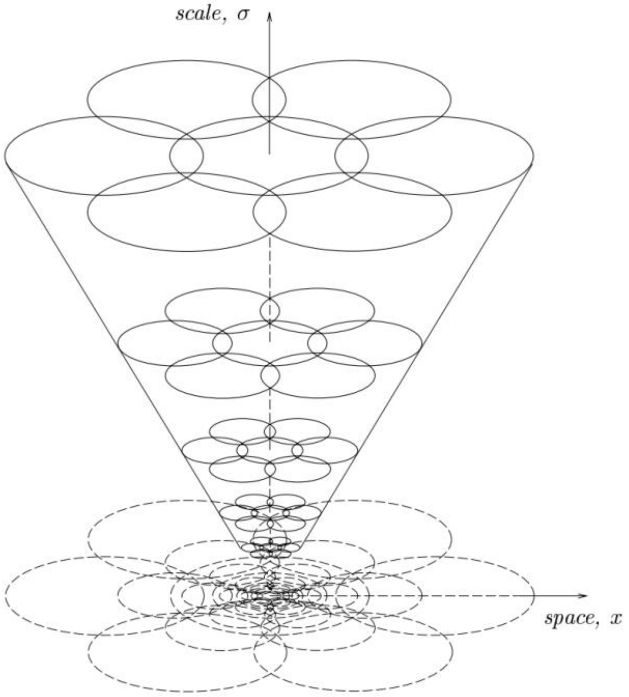


Figure 1-3 A schematic illustration of foveal system. It shows the receptive fields at different level. Note that the relative size of RFs is arbitrary. Image from (Lindeberg and Florack 1994).

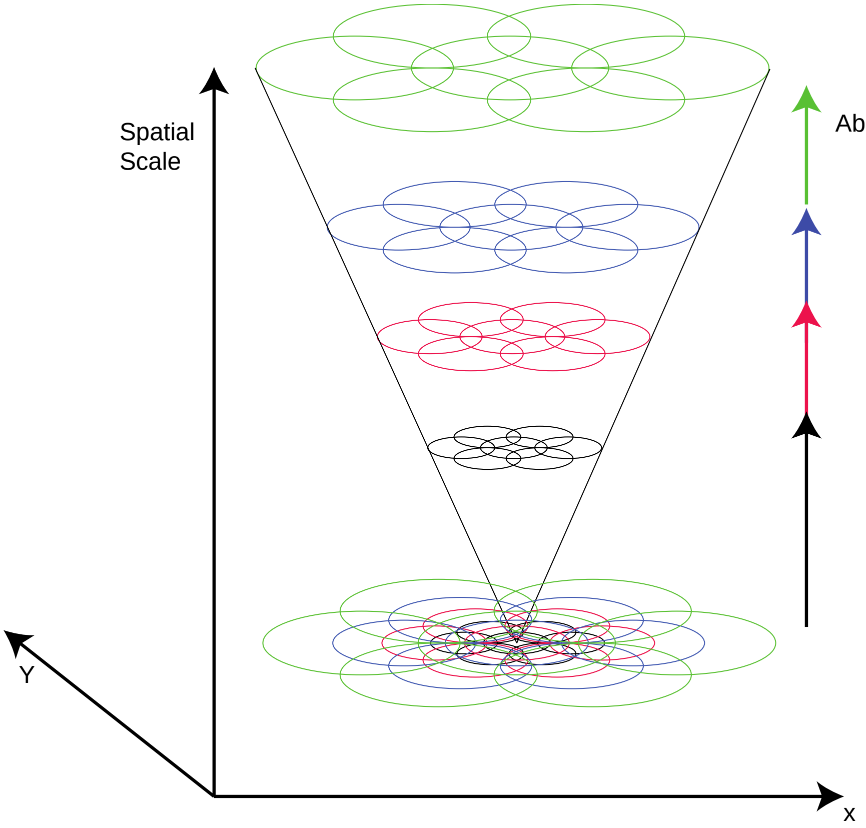


Figure 1-4 A schematic illustration of the concerned visual hierarchy. Colours are used to represent heterogeneous kernels at each scale.

1.2 Our Approach

In this work, we term the aforementioned recurrent connectivity conforming to the fast-brain hypothesis as *early recurrence*. To test the hypothesis that early recurrence improves early visual representation, we propose in this dissertation a computational model of early recurrence. The model formalizes computations of the two visual pathways of the primate visual system: a coarse-spatial-fine-temporal representation as the dorsal processing, and a fine-spatial representation as the ventral pathway. According to the fast-brain hypothesis, feed-forward processing in the dorsal pathway takes place before that in the ventral pathway. Via feedback connections, the dorsal representation modulates the processing in the ventral pathway. Figure 1-8 illustrates an example of the proposed computation in edge detection.

We further define a set of principled computational components. We firstly formulate feed-forward computations of a number of visual areas, including the retinal ganglion cells (RGCs), the LGN, visual areas V1, V2, MT and MST. Then we formulate the recurrent operation as a multiplicative inhibition. It is important to point out that our model is by no means to represent the true nature of the biology. However, to verify the proposed computation is consistent with biology, we applied the proposed computation to simulate two biological experiments: a figure-background segregation experiment (Hupé et al. 1998), and a Kanisza illusory rectangles experiment (Seghier et al. 2000b).

The essence of our work is to improve early-level visual representation by suppressing distracting features. Intuitively, the proposed computation has a wide range of applications in computer vision. One potential usage is in the self-driving vehicle, where the system needs to generate a visual representation of the surrounding traffic, and detect hazardous objects (e.g., pedestrians, other vehicles) based on this representation. In this example, early recurrence

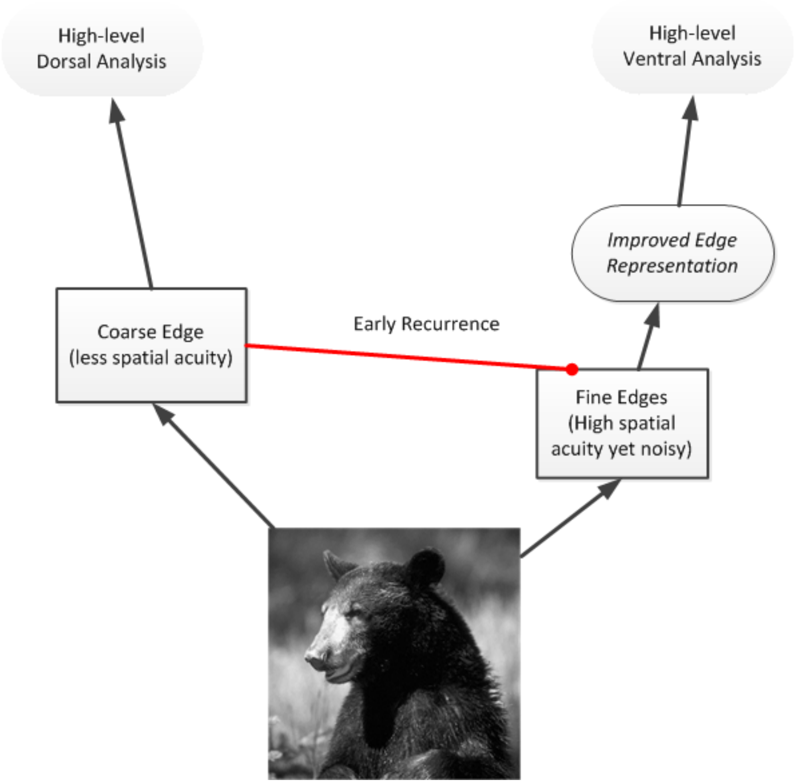


Figure 1-5 Apply early recurrence to improve edge detection. Input image is processed via the two visual pathways (black arrows). Early recurrence (red line) applies result of the dorsal processing to modulate ventral processing, leading to refined representation for further analysis.

may facilitate the on-board computer to improve the visual representation by suppressing irrelevant visual features.

To this end, we have applied our model in a variety of applications. In the application of visual saliency, three state-of-the-art saliency models have been implemented and revised to use modulated features to calculate the saliency representation. We conduct both empirical comparison and quantitative analysis with human fixation data. Results show that the modulated saliency representations significantly exceed their non-modulated versions. One of the modulated representations is further applied to background subtraction and scene recognition. Via quantitative comparison, we conclude that the proposed early recurrent modulation improves application performance in a robust and generally applicable manner.

In the application of edge detection, an early recurrent inhibition operator is developed based on our model. The operator is compared with another biologically-inspired operator (Grigorescu et al. 2003). The two operators are implemented following a standard edge detection paradigm. Using real images, we show that the proposed operator surpasses the competitor in most performance metrics. Our work is further used to improve the Canny edge detector (Canny 1986) and another edge detector (Martin et al. 2004). The results clearly show that the proposed model improve edge detection in real scenes.

1.3 Contributions

The main contributions of this dissertation are as follows:

1) A computational model of early recurrence.
2) A set of biologically inspired computational components and representations.
3) A computational framework to include the proposed computation components to model early recurrence. The framework has the following key characteristics:

a) A biologically plausible hierarchical framework to accommodate early recurrence at multiple hierarchical levels.
b) The framework works in a dynamic manner. The working order and temporal delays of early recurrent mechanisms are determined according to the actual primate visual system.

4) The proposed model simulates two biological experiments. Simulation results are consistent to the original studies.
5) Early recurrent modulation has been applied in different computer vision applications. Results show consistent improvements over prior-arts.

During the research, we have transformed some of these contributions into academic papers. They are listed as follows:

- Shi X, Wang B, Tsotsos JK, *Early Recurrence Improves Edge Detection,* BMVC 2013.
- Shi X, Tsotsos JK, *Background Subtraction via Early Recurrence in Dynamic Scenes,* ICPR 2012.
- Shi X, Bruce NDB, Tsotsos JK, *Biologically Motivated Local Contextual Modulation Improves Low-level Visual Feature Representations,* ICIAR 2012.
- Shi X, Tsotsos JK, *Improved Edge Representation via Early Recurrent Inhibition,* CRV 2012.
- Shi X, Bruce NDB, Tsotsos JK, *Fast, Recurrent, Attentional Modulation Improves Saliency Representation and Scene Recognition,* CVPR-Workshop 2011.

The rest of this dissertation is arranged as follows. Chapter 2 reviews the related theories and models. Chapter 3 proposes the computational model of early recurrence. To verify the proposed model, Chapter 4 simulates two biological experiments that support the fast-brain hypothesis. To show the influence of early recurrence to computer vision, we further investigate its usage in solving practical computer vision problems. Chapter 5 studies the application of visual saliency. Chapter 6 studies the application of edge detection. Last but not least, Chapter 7 provides the general conclusions.

Chapter 2. Literature Review

The ever-growing knowledge of the primate visual system continues to motivate new algorithms to improve performance of computer vision applications. However, the biological system is so complex, and there are so many gaps in our knowledge, such that making progress is usually non-trivial. We need to follow scientific methodologies to conduct research. One of the classic scientific methodologies follows the cycle of observation, hypothesis, prediction, experiment, and new theory; then new observation triggers the next round. For example, theories of the famous Yarbus fixation experiment have been evolved in such a manner. In the original work (Yarbus et al. 1967), the author proposed that deployment of eye fixation in the context of a task requires a top-down process. The observer's task could be predicted from his/her pattern of eye movement. For a long time, this idea had been well acknowledged. Recent experiments (Greene et al. 2012) challenged the conclusion by showing that subjects fail to identify the tasks performed by the observers based on the static scan paths (without seeing the order of fixations). However, more recent investigations (Borji and Itti 2014, Haji-Abolhassani and Clark 2014) overturned Greene's work and showed counter-examples to support Yarbus's original hypothesis. Of course, in this dissertation, we are not interested in this specific topic, but it is worth stating that the research path these debates following are important within the scientific process.

Let us consider the research on lower-level visual representations of the primate visual system. Since anatomical studies revealing its network structure, theories of visual processing from a connectionist perspective have emerged one after another. In general, these theories can be classified into two types: feed-forward theories and feedback theories. Feed-forward theories focus on explaining how the brain extracts and projects information from lower-level visual area to higher-level visual areas, while feedback theories emphasize the additional contribution of feedback or recurrent mechanisms. Since both feed-forward and feedback processing are embodied in the visual cortex, some elements of both types of theories have validity.

Motivated by recent feedback theories, we are interested in modelling the computation of lower-level recurrence, and investigating its utility in computer vision. Specifically, our biological foundation rests on two elements: lower-level feedback connectivity and timing of visual processing. This chapter will provide a literature review on important research that motivates our investigation.

In the following sections, we will first review the biological foundation of the primate vision, major feed-forward and feedback theories, timing of visual processing, and the hypothesis of early recurrence. The relationship between early recurrence and computer vision will be discussed. Specifically, we build a connection between early recurrence and the scale-space theory. The ultimate goal of this work is to show that early recurrence can fundamentally benefit computer vision applications. Thus, we will also review related computer vision systems.

2.1 An Overview of Biological Visual Systems

This section starts with a brief history of biological vision research and ends up with our up-to-date knowledge of the primate early visual hierarchy, relevant to this dissertation. A subset of the hierarchy is defined to contain the visual areas and connections needed to the current study.

The primate visual system is among the most complicated yet highly efficient information-processing networks in the world. It is a massive network containing a large number of feed-forward, feedback and lateral connections. Our ancestors began studying the brain from ancient times. However, the brain was not originally treated as an important organ. Ancient Egyptians thought the heart is the central unit that possesses intelligence. During the mummification process, the heart was the only organ left within the body (Wade et al. 2011). Our attempts to understand brain functions began during the 17th century BC (Wilkins 1992). Due to the limited technologies, until the 4th century BC, the brain was believed to merely function as a cooling system to bring down body temperature. Since the Hellenistic period, approximately the 3rd century BC, anatomical and physiological studies emerged. It was then realized that the brain indeed controls the nervous system. An important proposal during the later Roman Empire was that the brain controls the senses, and the nervous system controls the muscles. During the Renaissance, anatomical studies discovered that the brain contains a large number of components connected via nerves, and that each component has a specialized function (Van Laere 1993). At that time, researchers began to propose the brain structure and studied the brain in a region-by-region basis. The invention of microscope and staining procedure in the late 19th century opened the door to understand the brain from a neural network perspective. It was noted that different parts of the brain process sensations, capabilities and intelligence respectively. A large portion of the brain is recognized as dedicated to vision (Kandel et al. 2013).

Contemporary vision research started during the mid-20th century. In late 1950s and early 1960s, David H. Hubel and Torsten Wiesel announced several important observations that greatly expanded our knowledge of the visual system. By studying cats, they discovered the columnized neural arrangement in the primary visual cortex (Hubel and Wiesel 1959, Hubel and Wiesel 1962). In a later study (Hubel 1963), Hubel put forth a description of the visual hierarchy and the concept of the visual pathway, and further hypothesized how visual perception may arise. These works have deeply influenced later neuroscience and computational vision.

Brain studies on non-human primates via invasive cell-recording techniques revealed a complicated network structure. An early summary can be found in (Felleman and Van Essen 1991), see Figure 2-1. The primary visual cortex (area V1) stands at the lowest level of the brain hierarchy that receives most input signals from the retina. It has connections to many of the higher-level visual areas. Further, the authors listed the then-up-to-date connections among these areas, which imply far more complexity in possible processing strategies than a simple hierarchy. Based on the direction of signal projection, inter-neuron connections were categorized as feed-forward connections, feedback connections and lateral connections.

A recent update to Felleman and Van Essen's network (Lienhart and Maydt 2002) extends the notion of laminar distribution of neurons interconnecting visual areas with an index of hierarchical distance. It also confirms that the main visual pathways are mostly composed of short to medium distance connections.

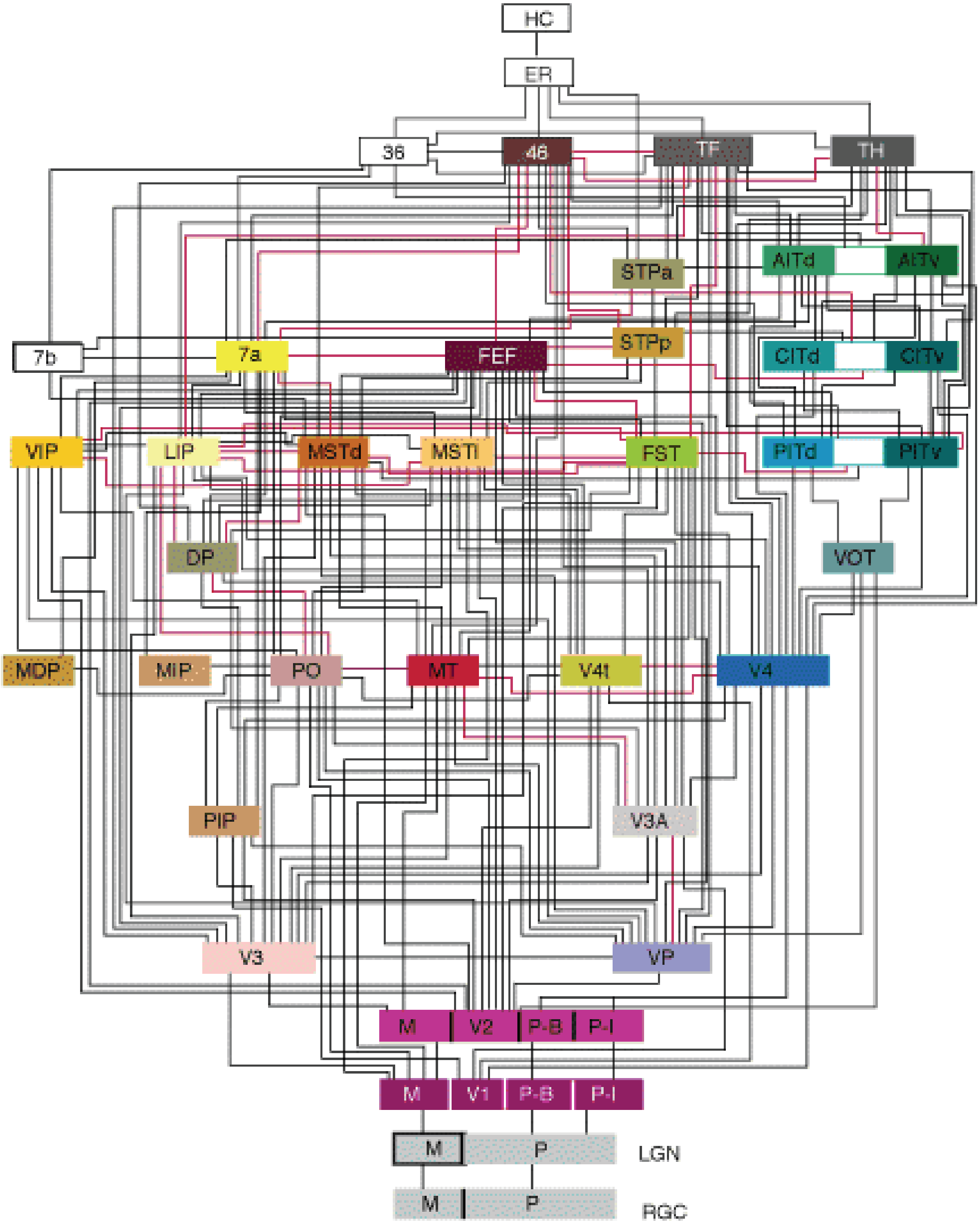


Figure 2-1 The hierarchy of visual areas and connections. Image from (Felleman and Van Essen 1991). A recent update (Lienhart and Maydt 2002) confirms that the main visual pathways ascending and descending the hierarchy have similar topographical structures.

Our investigation focuses on the early stage of the hierarchy, thus only a subset of visual areas and associated connections are within our scope, see Figure 2-2. In this view, visual processing starts from the Regina Ganglion Cell (RGC) through the lateral geniculate nucleus (LGN) into cortical areas. In humans and macaques, the LGN has three main types of cells. The magnocellular (M-) cell mainly responds to high-temporal frequency and low-spatial-frequency achromatic visual features. The parvocellular (P-) cell mainly responds to low-temporal frequency high-spatial-frequency chromatic features. The koniocellular (K-) cell has physiological properties between that of the P-cell and the M-cell (Xu et al. 2001). Our current work is mostly inspired by research on P- and M- cells, thus K-cells are not modelled at present. A large portion of the LGN connects to the primary visual cortex, V1 (Lamme et al. 1998, Casagrande and Xu 2004). Although there are other connections from sub-cortical regions to cortical regions (Sincich et al. 2004, Van Den Stock et al. 2011), they are irrelevant to the current study. From area V1, the visual hierarchy continues along two main visual pathways: the dorsal pathway and the ventral pathway (Mishkin and Ungerleider 1982, Norman 2002). The dorsal pathway stretches from area V1, via area MT, area MST, to the inferior parietal lobe. This pathway is mainly concerned with visuospatial processing (e.g. object localization, motion, spatial working memory, visually guided action and navigation, etc.). The ventral pathway connects the primary visual cortex to higher cortical structures to form object representations. The main visual processing areas include areas V1, V2, V4 and the inferior temporal (IT) areas. Each pathway is comprised of neurons that have distinct sensory selectivity, and often especially for the earlier layer, they may be considered as being specialized for particular visual features. How they may determine visual features in a manner that is invariant to illumination, viewpoint, and other external characteristics of image acquisition, and then how the pathway integrates this to form a stable view of the world are the major foci of the current research.

Classic theories of the ventral pathway gave rise to the idea that it is a serial hierarchy, with each sequential stage having progressively more complex selectivity and invariance than its lower stage. This sequential view excludes any feedback or lateral processing. Computation is strictly within the ventral pathway and proceeds in a feed-forward manner. However, recent studies suggest that the ventral pathway is actually a much more complicated recurrent network (Kravitz et al. 2013). Not only that, the ventral pathway has connections with the dorsal pathway at multiple stages. This suggests that, to reach an invariant visual representation, ventral processing is not isolated. However, whether feedback from the dorsal pathway is capable of impacting the ventral processing critically depends on the timing of visual processing between the two pathways.

Feedback connections have been shown important during visual processing. However, the effect of feedback differs from area to area, (Perkel et al. 1986, Van Essen et al. 1986, Henry et al. 1991, Salin and Bullier 1995, Martinez-Conde et al. 1999). Interestingly, most feed-forward and feedback connections within the early visual hierarchy show a paired pattern (Battaglini et al. 1982, Bullier et al. 1984, Ferrer et al. 1988, Ferrer et al. 1992, Salin et al. 1992, Lienhart and Maydt 2002).

An important property to characterize a neuron is by its receptive field (RF). Specifically, the classical definition of receptive field refers to the spatial region where a change of stimulus causes a response of the neuron. Neurons in the higher-level visual areas have larger RFs. That is, a higher-level neuron will respond to stimuli or visual representation over large spatial field. However, with feedback mechanisms, the definition of RF extends beyond this

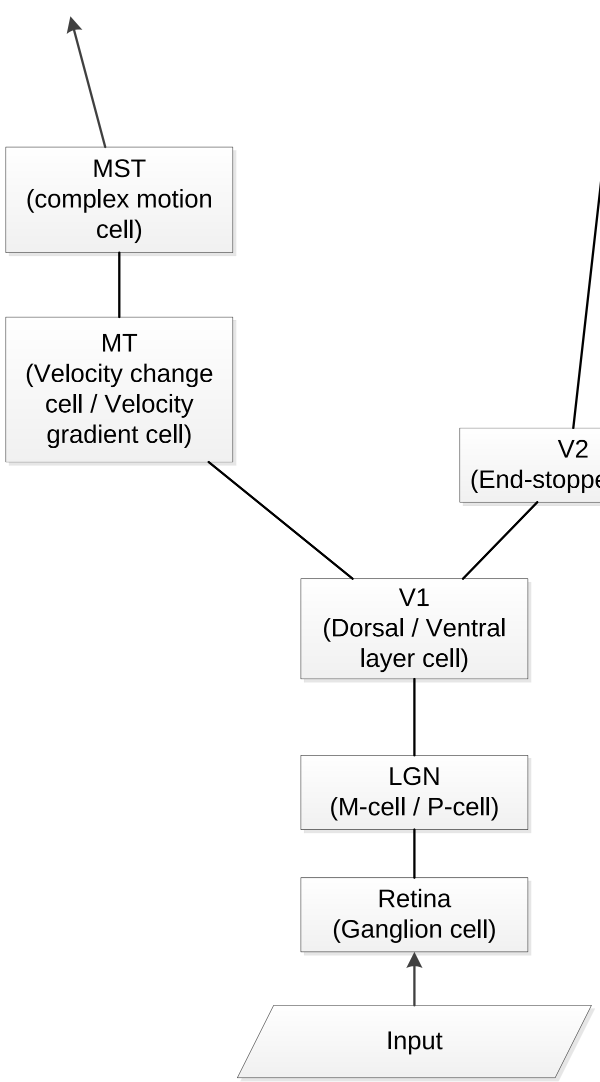


Figure 2-2 The current work focuses on a subset of the primate visual hierarchy, where visual processing starts from the Retina, through the LGN into area V1. From area V1, the visual hierarchy continues along two main visual pathways: the dorsal pathway (including area MT and area MST) and the ventral pathway (including area V2). Texts in the brackets indicate the correspondent cell types that the current work simulates.

classical sense (Angelucci and Bullier 2003). A non-classical RF describes the extended region that, although it cannot drive neural response directly, is capable of exerting suppressive or facilitative effects on the response to the presentation of stimuli in the classical RF. For example, by reverse correlation, it has been reported that the classical and non-classical RF of V1 neurons are 0.45° and 1°.

It is suggested that feed-forward connectivity determines RF properties and transforms visual input into behavioural responses, and feedback connectivity together with lateral connectivity plays the roles that tune RF properties based on higher-order perception, attention, and visual awareness (Lamme et al. 1998, Angelucci and Bullier 2003, Angelucci and Bressloff 2006). Many functional role studies have revealed the influence of feedback over area V1 and area V2 (Sandell and Schiller 1982, Alonso et al. 1993, Martinez-Conde et al. 1999, Angelucci et al. 2002, Angelucci and Bressloff 2006), and over higher-level visual areas, such as area MT (Hupé et al. 1998), area MST (Berzhanskaya et al. 2007), and area V4 (Ungerleider et al. 2008).

The current study particularly focuses on the lower-level recurrent connections that cross the two visual pathways. However, before reviewing the details, it is necessary to change our topic to computational neuroscience. In the next section, we will review how existing theories put forth feed-forward and feedback processing, and associated implications to computer vision.

2.2 Computational Models of Feed-forward and Feedback Processing

The goal of computationally modelling the primate visual system is to quantitatively explain certain aspects of visual processing. Connectionism is a common approach to model neural network processing (Feldman and Ballard 1982). A connectionist model usually starts from building a hierarchical network of the visual system, based on which the model specifies the computation at each hierarchical layer.

Based on type of modelled connectivity, one can categorize existing models of the primary visual system into feed-forward models and models with feedback. In what follows, we will first review a number of pure feed-forward models, which do not contain feedback components. The review focuses on how these models calculate invariant visual representations in pure stimulus-driven fashions.

2.2.1 Pure Feed-forward Models

Inspired by Hubel and Wiesel's discovery of the visual hierarchy and neural response patterns, many models start from simulating neurophysiological properties. Marr's well-known work is a classic example (Marr 1982). The initial representation is a two-dimensional visual array (retinotopically aligned to visual fields) derived from the retina. This representation then projects to higher levels. The feed-forward process continues until a three-dimensional interpretation of the scene is fully reconstructed. Specifically, the process contains three stages:

- **2D Primal Sketch**: to build a 2D primal sketch based on extracted visual features (zero-crossings, edges and curves).
- **2.5D Sketch**: to construct local surface orientations, contours, discontinuities, and depth information based on the 2D primal sketch.
- **3D Sketch**: to construct shapes and their spatial relationships on top of the 2.5D sketch.

Based on the then-up-to-date knowledge of the primate visual hierarchy, visual processing was proposed as a purely feed-forward fashion. Figure 2-3 illustrates the concept from an information-flow perspective, which illustrates the basic information flows of the visual hierarchy. Marr did not specify the anatomical localizations that compute these representations. However, in his theory, there is no feedback mechanism at any stage of the hierarchy. Without feedback tunings, the visual processing is hardwired, i.e., the processing hierarchy is a static structure. It is now known that the primate visual hierarchy is plastic, and can not only learn over time but is also dynamic and adaptable to the input stimulus and visual task (Kourtzi and Dicarlo 2006). The view of the static visual process is now discarded; however, the legacy remains influential.

In (Fukushima 1980, Fukushima 1988), a hierarchical multi-layered neural network, the so-called Neocognitron, was proposed that is capable of performing position-invariant visual pattern recognition. Specifically, position-invariance refers to the ability to recognize a pattern regardless of where it is found in the visual scene. Neocognitron model includes an input layer and a number of simple cell / complex cell (S-cell / C-cell) pairs that are inspired from the biological vision, e.g. (Hubel and Wiesel 1959). In the model, recognizing an object is a recursive process of successive S-cell / C-cell refinements. The basic routine is that the S-cells extract visual features and the C-cells combine responses of the S-cells over different positions, thus achieving localized position-invariance. The highest-level C-cells compute the ultimate representation, which is then used for the self-organized learning system for recognition tasks (Figure 2-4).

One disadvantage of the feed-forward Neocognitron network is that the performance drops when two or more patterns present simultaneously. To solve this problem, a top-down propagation process was added to the conventional Neocognitron network in later extensions,

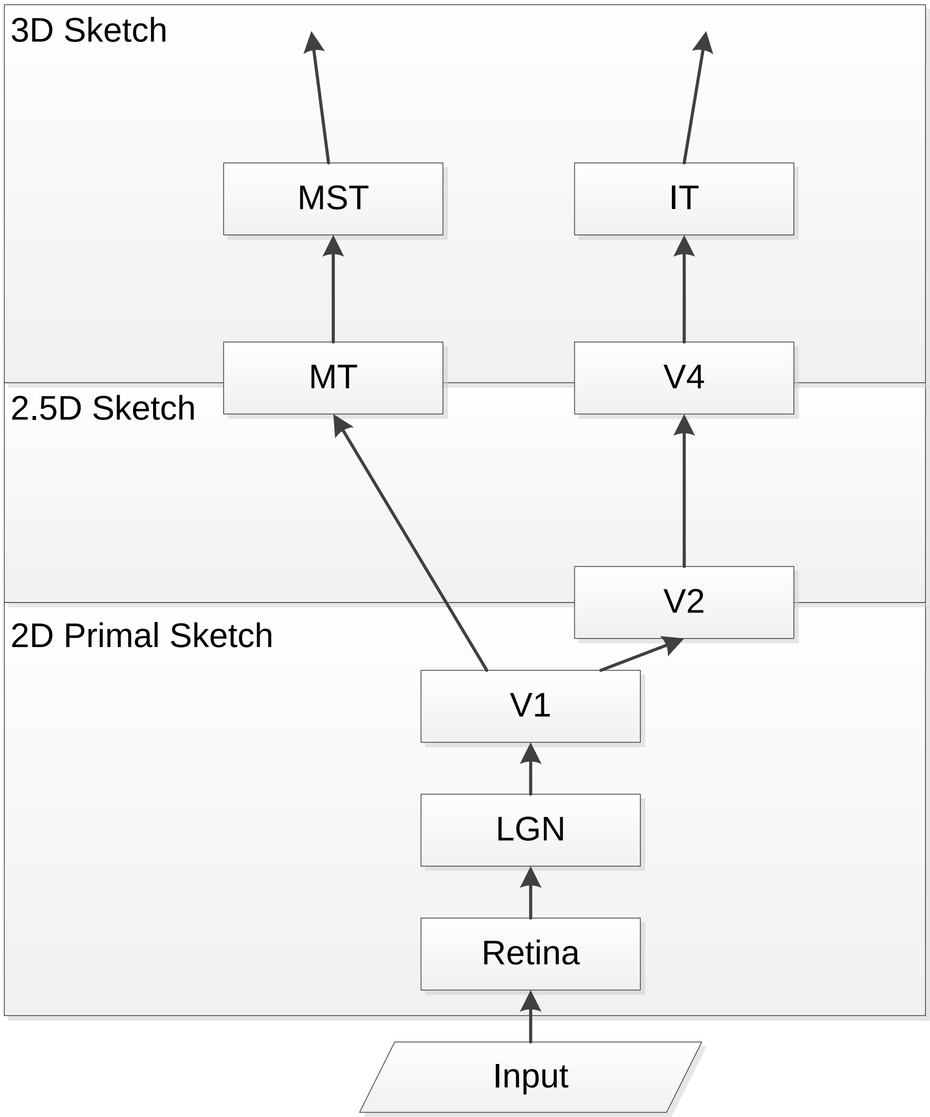


Figure 2-3 Mapping the visual hierarchy with the 3-stage vision expressed by Marr. The visual hierarchy works in a pure feed-forward fashion. Input stimuli are captured at the lowest level (Retina). Upper level receives input from lower level to form complex visual representation (from 2D primal sketch to 2.5D sketch). Finally, the top layer generates the 3D representation to restore the scene.

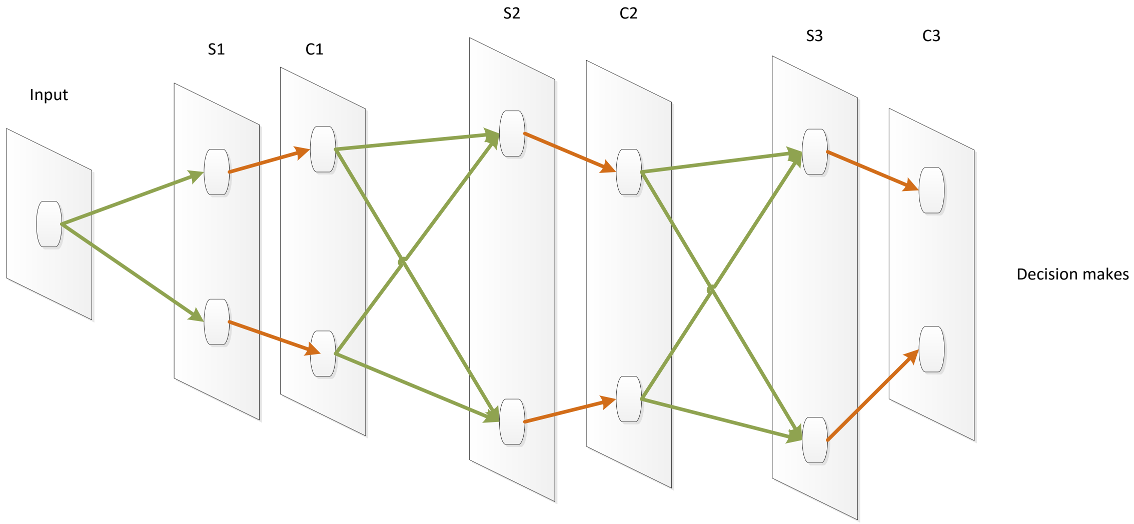


Figure 2-4 The typical architecture of the Neocognitron (Fukushima 1980, Fukushima 1988). It includes an input layer and a number of S-cell / C-cell pairs. The basic routine is that the S-cells extract features (green arrows), and the C-cells combine responses of the S-cells over different positions (red arrows) to achieve position invariance. Output of the highest-level C-cells is then used for the self-organized learning system for recognition tasks. Note that in this figure, information flow is feed-forward only.

see (Fukushima 2013) for a review. In an early proposal (Fukushima 1986), the top-down processing (called the selective attention model) facilitates the network to segment and recognize targets one after another. In recent revisions (Fukushima 2001, Fukushima 2005), the network is further improved to recover occluded object parts. To do this, the network first detects the occluded object; a mask layer is defined to suppress response that is irrelevant to the occluded pattern. As such, by the top-down suppression, the network has the ability, not only to recognize occluded patterns correctly, but also to restore the occluded parts.

Neocognitron has been successfully used in many hand-writing recognition applications (Fukushima 2007). In these applications, input images are mostly hand-writing samples. However, for generalized object recognition, how to compute invariant feature representation is still an on-going major challenge.

Poggio and his colleagues faced this challenge with a two-step framework (Poggio et al. 1988). In the first step, different feature extraction algorithms are used. Outputs of these algorithms are discontinuity representations with regards to specific features. Each discontinuity representation is a retinotopic intensity map, where a higher intensity value indicates a higher probability of feature presenting. In the second step, the computed feature maps are combined into a surface discontinuity map, which is further used for object recognition.

In (Riesenhuber and Poggio 1999), the authors proposed a computational hierarchy of object vision. This work is claimed to be consistent with the processing in the primate ventral pathway. Specifically they modelled the processing of area V1 and the inferotemporal cortex (IT), which accounts for many recognition tasks. The hierarchy shares many common characteristics with Neocognitron. For example, they both have layers of S-cell and C-cell pairs to successively achieve position-invariant. An HMAX operation was proposed for feature combination, which includes two methods: sum of features and max of features. The sum of features is a weighted linear summation over different feature maps that would explain the increase in complexity of the optimal stimulus driving cells for object recognition. The max of features is a non-linear operation that pools over slightly distorted versions of the same feature-set to provide the substrate for building invariant representations. Later, the model was refined (Serre and Riesenhuber 2004) to couple with biologically-inspired filters for S-cells, and tuning properties for C-cells.

Another feed-forward approach is by statistical analysis. One successful example is the SIFT descriptor (Lowe 1999, Lowe 2004). In order to compute local scale-invariant features, the algorithm starts from transforming the input image into a scale-space representation. The author proposed to use Difference-of-Gaussians filter at different spatial scales to extract image intensity variations, based on which extrema in both spatial and scale neighbours define a set of key locations (descriptors). Each key location is assigned a canonical gradient-based orientation so as to describe the key location invariantly to rotation. The image is then characterized by these key locations. To perform object recognition, a typical method is by descriptor matching (Skrypnyk and Lowe 2004, Brown and Lowe 2007). Similar statistical based feature descriptors have been recently proposed and successfully used in various machine learning applications (Dalal and Triggs 2005, Vedaldi and Fulkerson 2010).

Progress on feed-forward artificial neural networks (ANN) triggered a series of applications that address machine learning problems. Among them, the Convolutional Neural Network (CNN) is perhaps the most relevant example. LeCun and his colleagues proposed a back-propagation learning network for character recognition (Lecun et al. 1990). The hierarchy of the network simulates S-cell and C-cell pairs. S-cell operation is equivalent to a convolution (thus called convolution network), followed by an additive bias function and a sigmoid function. The weights to these functions are trainable. C-cell operation performs local-averaging and sub-sampling. By such a paired operation, the network achieves location and distortion invariance. The output layer measures the Euclidean distance between computed vector (representing input) and parameter vector (representing a class). Back propagation is applied to train the network. The CNN was then developed with a new learning paradigm (Lecun et al. 1998), called Graph Transformer Networks, for more generalized hand-writing recognition.

Recent advances of ANN include the popularized Deep Learning (Lecun et al. 2015). It intends to learn representations of data by using hierarchical architecture (similar to the visual hierarchy), with complex structures and multiple non-linear transformations. Deep Learning has significantly improved the state-of-the-art in speech recognition, visual object recognition, object detection and many other domains.

2.2.2 Feedback: A Missing Component

Feedback is an important part of visual processing that did not appear in Marr’s three-stage vision hierarchy. We now know that the primate visual system includes a massive feedback network. It is our goal to contribute to an understanding of the role of feedback in visual processing. Numerous other studies have the same goal. For example, it has been suggested that our brains use feedback to adapt visual processing to current task demands (Jones et al. 1997, Baluch and Itti 2010, Greene et al. 2012), past experience (Lee and Mumford 2003, Kersten et al. 2004, Borji et al. 2012) and scene context (Chun and Jiang

1998, Bar 2004, Oliva and Torralba 2007). In (Lamme and Roelfsema 2000, Roelfsema 2006), the authors presented neurophysiological evidence and models that connect feedback to perceptual grouping tasks. In a more recently study (Felzenszwalb et al. 2010), the authors presented convincing evidence suggesting that the cortical processing is not purely hierarchical and feed-forward. They claimed that in order to know how the visual system processes fine-grained information at a particular location, it is necessary to integrate information about the surrounding context over the entire visual field via feedback mechanisms. Grouping and segmentation are crucial to understanding vision, and must be understood on a global scale.

2.2.3 Feedback Theories

It is commonly believed that feedback control in vision facilitates the feed-forward processing via different types of neural modulations. The literature suggests that feedback modulation can impact feed-forward processing by at least three aspects: strengthening feed-forward response (Haenny et al. 1988, Motter 1993, Luck et al. 1997), attenuating feed-forward response (Reynolds et al. 1999) and enhancing baseline activity (Luck et al. 1997, Kastner et al. 1999).

To model a feedback mechanism, we must answer three questions: where does the feedback come from? What is the feedback representation? And how does this representation affect visual processing? We know that feedback takes place at different levels of the visual hierarchy: attentive feedback, intermediate feedback, and early recurrent feedback. These fall within the broad category of attentive processes, following the definition in (Tsotsos 2011):

"Attention is the process by which the brain controls and tunes information processing."

Attentive feedback has been actively studied for decades. However, due to its complicated nature, the exact neural origin of attentive feedback is not clear. In many theories, it is believed that visual attention is generated outside the visual hierarchy, i.e., in much higher-level cortical areas (Desimone and Duncan 1995, Kastner and Ungerleider 2000, Gregoriou et al. 2009). The representations are abstract, yet capable of containing information about where and what to look.

Among existing attention theories, an early conceptual and influential model was the Feature Integration Theory (Treisman and Gelade 1980a). Its computational proposal was later proposed in (Koch and Ullman 1985). In this model, a topographical central representation, or the so-called saliency map, is computed in pure feed-forward fashion based on combinations of extracted visual features. A Winner-Take-All (WTA) procedure is developed to compute the most salient region or feature from the central representation as attention. Many models employ this feature-driven process and the WTA computation (Phaf et al. 1990, Ahmad 1991, Itti et al. 1998, Bruce and Tsotsos 2009a).

Another well-known feedback model was proposed by Grossberg and colleagues, the Adaptive Resonance Theory (ART) (Grossberg 1987, Carpenter and Grossberg 1990, Carpenter et al. 1991). ART is a self-organizing neural network for pattern recognition in response to arbitrary sequences of input patterns. The system involves both feed-forward feature computation and feedback expectation modulation. Feedback takes form as priming of expectation, or learned prototype vectors. In this model, the feed-forward representation is compared with the feedback representation to measure the belongingness of features to a known pattern.

In (Anderson and Van Essen 1987), the primate visual system has been described as a set of shift-able circuits, or the so-called Shifter Circuits, which provide a generalized routing strategy for dynamic control of information flow to calculate visual attention (in the Koch and Ullman sense). The strategy can be applied to a broad range of visual tasks, such as stereopsis, motion analysis and visual attention. The model includes a bottom-up stimulus-driven saliency structure to control the routing of neural arrays between the LGN and area V1. Shifting of focus is accomplished through feedback controls, during which ascending pathways are selectively suppressed by inhibitory neurons (through macro-shifting circuitry). However, as a general framework, Shifter Circuits do not include processes for modulating neural processing other than selection of inputs.

Tsotsos and colleagues proposed the Selective Tuning (ST) hypothesis (Tsotsos et al. 1995). Its earlier prototype (Culhane and Tsotsos 1992) introduced a routing strategy and the concept of inhibitory beam that computes a selection based on both feed-forward and feedback processing. The up-to-date version of the model simulates a variety of attentive mechanisms, several involving feedback. Among these is a top-down biasing of the visual processing hierarchy based on task demands, a recurrent localization process that traces neural connections from top to bottom of the hierarchy in order to select the stimulus location of the most strongly responding neurons at the top, and a top-down suppression mechanism that is used to reduce interference among competing stimuli on feed-forward pathways. ST has predicted new characteristics of biological visual processes that are supported by experiments on human and non-human primates (Tsotsos 2011). Several more specific aspects of ST have also appeared, such as motion selection (Tsotsos et al. 2005), active search (Zaharescu et al. 2005), and features binding (Tsotsos et al. 2008).

As for modelling intermediate feedback and early recurrent feedback, the literature is limited despite the fact that there are plenty of psychophysical and neurophysiological studies on lower-level feedback mechanisms. It is thus our intention in this work to provide one way to model early recurrent feedback.

2.3 Timing of Visual Processing and Early Recurrence

Now we continue our review on the topic of feedback mechanisms. In this work, timing is a critical factor in driving early recurrence. As discussed in previous sections, the LGN contains three major types of cells (Irvin et al. 1986, Maunsell et al. 1999): magnocellular (M-) cell, parvocellular (P-) cell, and koniocellular (K-) cell. In this work, M and P cells will be modelled. Table 2-1 provides a comparison. A key notion is that the M-cells have a faster response time to input than that of the P-cells (Lindeberg 1998). If we couple this timing property with the spatiotemporal frequency that they respond to, we see that the visual processing of low-spatial-and-high-temporal frequency band (by the M-cells) is performed prior to the visual processing of high-spatial-and-low-temporal frequency band (by the P-cells).

Starting from the LGN, Bullier and his colleagues studied the timing of visual processing by examining experimental data of their own and other groups (Bullier 2001). They concluded that the latencies of the input signal reaching different visual processing areas do not conform to a nice feed-forward hierarchical pattern in the Hubel and Wiesel sense. Table 2-2 briefly summarizes response latencies among several visual areas.

In a similar study (Schmolesky et al. 1998), responding latencies evoked by flashing stimuli were measured in the LGN and in a number of cortical areas in anesthetised macaques. In the LGN, it is observed that the magnocellular cells have a response time that is an average of 17 milliseconds earlier than the parvocellular cells. Visual responses occurred in area V1 before any other cortical areas. The next wave of response occurs concurrently in areas V3, MT, MST, and FEF. Visual response latencies in area V2 and V4 were progressively later and more broadly distributed (Figure 2-5).

Figure 2-6 overlays some of the temporal response properties just described to a hierarchical view of the areas of Table 2-2. It shows that visual areas in the dorsal pathway respond to visual stimuli with very short delays compared with visual areas in the ventral pathway. Dorsal V1 responds to LGN M-cell input at about 40 milliseconds after stimuli onset. Results are sent via the dorsal connections to area MT, MST, FEF and 7a. The top layer of the dorsal hierarchy (area 7a) has a mean response delay of 90 milliseconds. However, at this moment, the same visual input is being processed in area V2 along the ventral pathway, which is a significantly lower-order visual area compared with area 7a. Further, it takes a longer time for the top layer of the ventral pathway (e.g., area IT) to respond to the visual stimuli. The fact that these visual areas respond to feed-forward stimuli at different temporal delays suggests that the two visual pathways are distinguished from each other by more than just physical connectivity and spatial receptive field patterns.

In efforts to understand the rationale behind such asynchronized information routing, researchers have proposed that the primary visual system may include feedback mechanisms that use results of early dorsal processing to modulate ventral processing. In (Hupé et al. 1998, Bullier 2001), the authors showed that feedback from area MT plays a role in differentiating a moving or flashing figure from background pixels, particularly for low saliency stimuli. They proposed that the neural response to visual stimuli at area V1 is modulated by feedback from area MT. They further hypothesized that recurrence acts in a

Table 2-1 LGN neural receptive field characteristics and connectivity

LGN type	Receptive field size	Corresponding Photoreceptor	From Retinal Ganglion Cell	To Area V1	Spatiotemporal frequency and Chromatic Sensitivity
Magnocellular	Large	Rods	Parasol Cells	Dorsal layers	Low spatial High temporal Achromatic
Parvocellular	Small	Cones	Midget Cells	Ventral layers	Band spatial Low temporal Chromatic

non-linear fashion to improve the gain of the centre mechanism and that it combine with horizontal connections to generate the centre-surround interactions.

In (Hupé et al. 2001), the authors investigated the time course of the recurrence. In the experiment, area MT neurons were inactivated. They showed that the response time of the V1 neurons was significantly affected by the inactivation. For the majority of the neurons marked in area V1, V2 and V3, the response decreased. Similar observation was reported with flashing stimuli. In both experiments, response latency was measured, which indicated that neurons in area V1 were affected with the earliest time by the feedback from area MT. These results indicated that recurrence from the dorsal pathway influences the ventral processing in very low levels.

As a result of these observations, Bullier developed his integrated model of visual processing (Bullier 2001). In the model, information from the magnocellular layers of the LGN through the dorsal pathway is very rapidly communicated. Results from this “first-pass computation” then projects back via recurrent connections to the ventral layers of area V1, and actively modulate the ventral processing. In Bullier’s view, the first-pass computation modulates ventral processing all the way to the Inferotemporal cortex.

To sum up the work, they concluded the areas within the dorsal pathway that respond to input stimuli earlier than those in the ventral pathway. Further, they concluded that there are rich recurrent connections across the two pathways especially from higher-level dorsal regions to lower-level ventral regions. Following these observations, they proposed the so-called fast-brain hypothesis: if timing differences permit, then the visual system might utilize these early recurrent connections to modulate ventral processing. Figure 2-7 illustrates this idea.

Table 2-2 Response latencies among early visual areas. Response latencies among early visual areas. Earliest and median latencies are recorded from monkeys in different studies. Earliest latencies refer to the delays observed with 10 percentile neural activations. At a brief comparison, it shows that the dorsal pathway requires shorter time than the ventral pathway to process visual input.

Visual Area	**The Dorsal pathway**					**The Ventral pathway**			
	V1	**MT**	**MST**	**FEF**	**7a**	**V1**	**V2**	**V4**	**IT**
Earliest latency (millisecond)	20-31	25	35	35 - 45	50	40	55	100	100
Mean latency (millisecond)	40	45	45	65	90	65	85	100 - 150	100 - 400

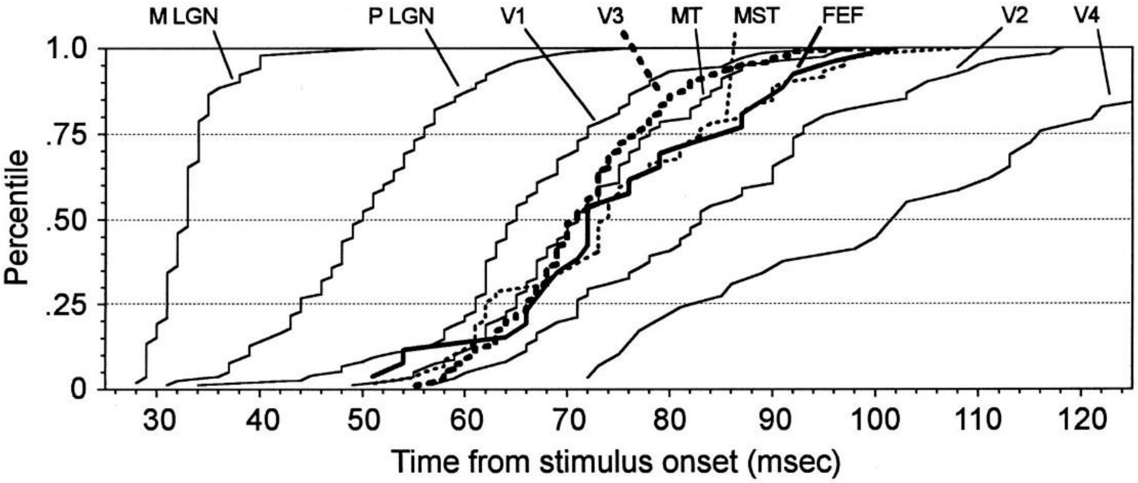


Figure 2-5 Timing of visual processing in various visual areas observed by (Schmolesky et al. 1998). The data observed are consistent with those reported by Bullier and colleagues, that the M-cells and the dorsal pathway conducts signal faster than the P-cells and the ventral pathway.

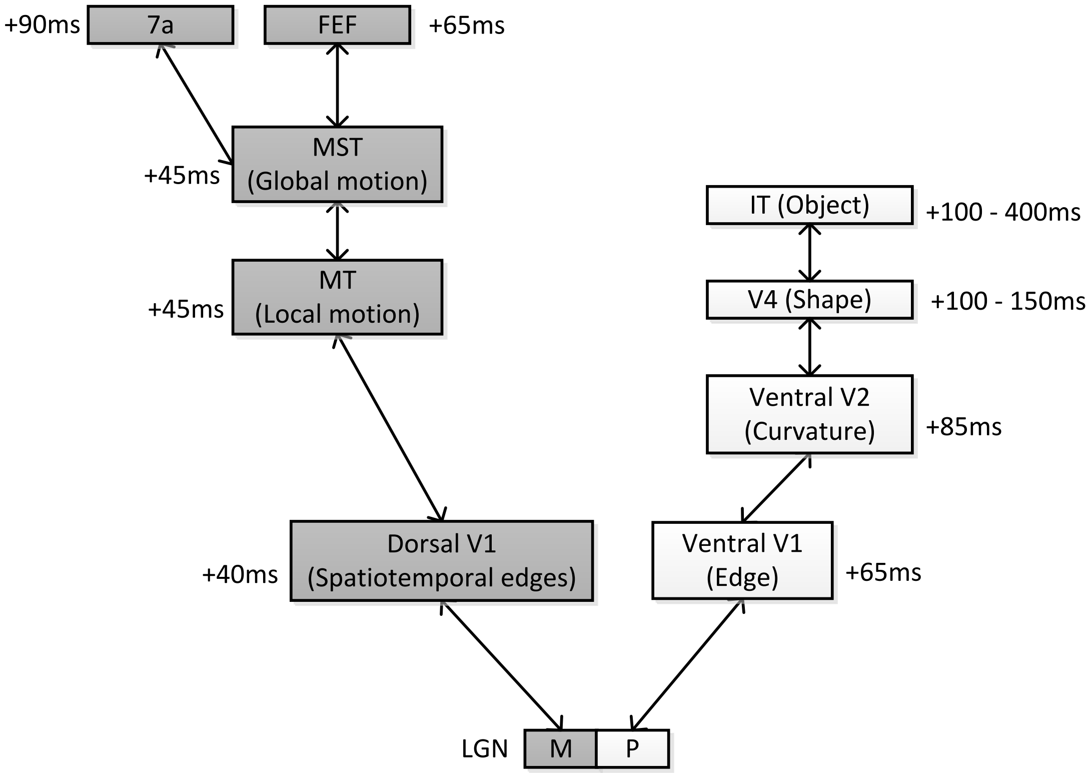


Figure 2-6 Response latencies of the visual hierarchy. Darker boxes indicate the M-cells and visual areas in the dorsal pathway. Lighter boxes indicate the P-cells and visual areas in the ventral pathway. Texts in brackets show visual features extracted. Numbers besides each box indicate the mean responding latencies as summarized in Table 2-2.

The functional roles of early recurrence have been studied. For example, a figure-ground discrimination experiment (Hupé et al. 1998) showed that dorsal feedback increases the difference between neural responses to a bar moving on a stationary background and those to the same bar moving together with the background. An fMRI study (Seghier et al. 2000a) using moving Kanisza illusory rectangles found strong and reproducible signals in area V1 and area V2 caused by area MT recurrence. The authors noted that MT-V1 feedback selectively strengthens the boundary signals in the chopstick illusion experiment to trigger boundary completion and figure-ground separation. It is further suggested (Bayerl and Neumann 2006) that MT-V1 modulation generates localized motion estimation, which is impossible by pure feed-forward interactions.

Although evidence suggests that impacts of early recurrence widely exist in the primate visual system (Rauss et al. 2011), attempts to model such mechanism have not had much progress since Bullier's early efforts (Nowak et al. 1997, Hupé et al. 1998). In the big picture of vision, little is known about the significance of early recurrence.

In computer vision, a related fast magnocellular mechanism has been modelled as context modulation to facilitate object recognition. Visual context may provide constraints to influence visual processing. Torralba and colleagues (Torralba 2003, Oliva and Torralba 2007) modelled one aspect of visual context from a scene perspective. Context is modelled as a topographical representation (scene Gist) of the structural summary of an input scene. Based on learned knowledge, a scene Gist is used to estimate the likelihood of a scene containing a target, and it further predicts the probable location/scale of the target. A Gist representation serves to facilitate object recognition, where it primes the input image with a multiplicative operation that suppresses irrelevant image regions.

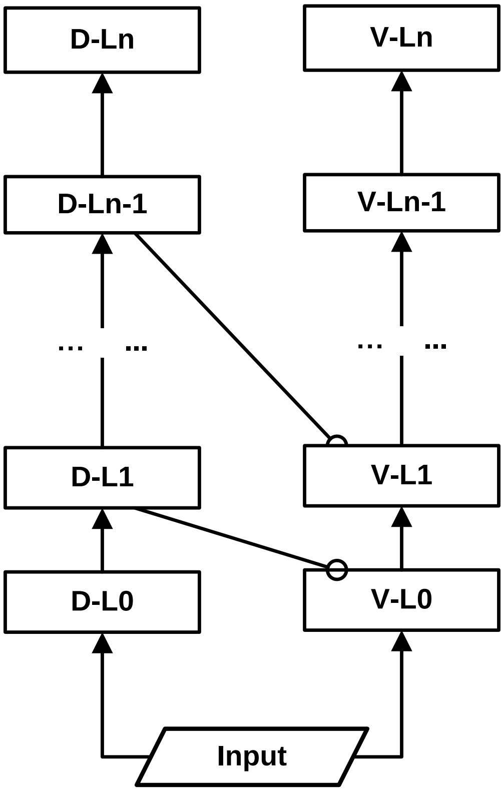


Figure 2-7 Early recurrence feedback paradigm. The figure assumes visual stimuli project through two visual pathways separately, with each pathway computing different visual features and transmitting information at different speeds. D-Li represents layer i of the dorsal pathway. V-Li represents layer i of the ventral pathway. In this example, early recurrence is between the dorsal area (D-L1) to the ventral area (V-L0), and between D-Ln-1 and V-L1, respectively.

A similar concept (Bar 2004) is proposed with more neurophysiological foundations. Using functional imaging combined with MEG recording, Bar and colleagues found that the main sites in the brain to mediate both spatial and non-spatial context are near the parahippocampal cortex and retrosplenial cortex. Context is activated as early as 130 milliseconds after stimuli onset. In their model of contextual facilitation, context is represented using context frames. Rapid bottom-up information through the magnocellular pathway activates a context frame, which then feeds back to IT. Intersections of context frame and object perception in IT yield a unique object.

2.4 Connections to Scale-Space Analysis

Another motivating component for this research is the scale-space theory (Witkin 1983). The goal of scale-space analysis is, by studying input signal at multiple scales, to create a scale-invariant visual representation for high level visual processing. The main motivation behind this is that a target represented in an image manifests itself as a meaningful entity only over certain ranges of spatial scale. Without a-priori information, an automated image processing system should have the adaptability to process objects of different scales.

The basic approach of scale-space analysis is to convolve an image with a group of Gaussian kernels with different kernel variance. As the variance increases, the convolved image becomes coarser, with high-frequency pixel variations (i.e., image noise and high-frequency edges) gradually disappearing. Stacking the convolved images by sorted by variance constructs a representation in the scale space. This representation has the desired character of causality, where no spurious edge should be detected as the convolved image become coarser. In order to detect edges, in (Witkin 1983), the author proposed a covering algorithm to compute a stable covering across the scale space. It was shown that output with stable variance has a marked correspondence with perceptually salient object edges. However, as Gaussian variance increases, the output representation loses spatial accuracy. The edge detected in the stable variance presentation may be severely shifted from the actual edge (in the original image). To cope with this loss of spatial accuracy problem, anisotropic diffusion has been proposed to preserve location information during the construction of scale-space representation (Gregoriou et al. 2009).

Figure 2-8 illustrates a one-dimensional example. Image (A) shows the Gaussian envelope at different variances. Image (B) shows the result of applying Gaussian kernels progressively over the original signal. Clearly from image (C), signals are successively smoother, yet are still able to catch the intuitive notion of fine-scale information.

The idea can be extended to two-dimensional image processing. During the 1990s, Lindeberg and his colleagues conducted research that applied scale-space analysis to solve computer vision problems. In his PhD dissertation (Lindeberg 1991) and later in (Lindeberg 1991, Lindeberg 1993), he presented a method to transform an image into the discrete scale-space domain. In the scale-space hierarchy, the lowest level represents the original pixels. Higher-level representations are computed by convolving the image with Gaussian filters of different variance kernels. This hierarchy is used for two-dimensional blob detection, which is similar to the 1D stable-cover algorithm (Witkin 1983).

Lindeberg’s blob representation has been used to extract image structure. The author has used this model to solve a number of computer vision applications, including perceptual saliency, edge detection, histogram analysis, and junction classification. In later studies (Lindeberg 1994, Lindeberg 1998), the author proposed a strategy of automatic scale selection. By detecting local extrema across scales, it complements classical scale-space theory by providing an explicit mechanism to predict the most informative scale.

The relationship between the scale-space theory and the primate visual system was discussed in (Lindeberg and Florack 1994). The work provides an idealized model based on first principles. A foveal system is defined with a dense distribution of receptive fields analogous to neurons in the primate visual hierarchy over space and scales (Figure 2-9).

There is one aspect of the biological visual hierarchy that seems inconsistent with scale-space analysis. In scale-space, the only parameter that changes as scale increases is the kernel size; that is, the nature of what each kernel computes does not change. For example, each kernel detects edges of the same kind but at different scales. However in the primate visual hierarchy, as processing level increases, not only does the scale change but also the nature of the information processing. For example, the primary visual cortex computes local translation patterns, while the MST computes global motion patterns. This makes the direct application of classical scale-space methods a bit problematic.

A conceptual connection may be drawn between the fast-brain idea and the scale-space theory. We note that both concepts focus on the lower-level visual areas. In both theories, the computation starts from basic feature processing. In the following chapters, we will propose and show that early recurrence may be playing a role to automatically adapt visual processing to the correct scale.

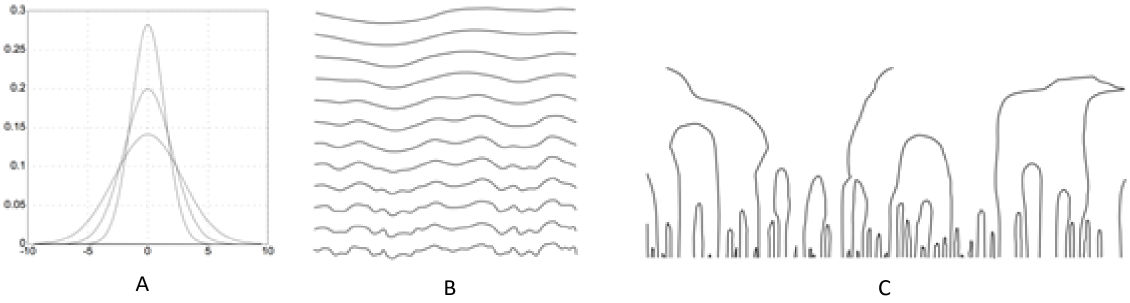


Figure 2-8 (A) One-dimensional Gaussian kernels with different standard deviation. (B) A scale-space representation is to generate a one-parameter family of derived signals in which the fine-scale information is successively suppressed. (C) Since new zero-crossings cannot be created by the diffusion equation in the one-dimensional case, the trajectories of zero-crossings in scale-space (here, zero-crossings of the second derivative) form paths across scales that are never closed from below. Images are from (Lindeberg 1994).

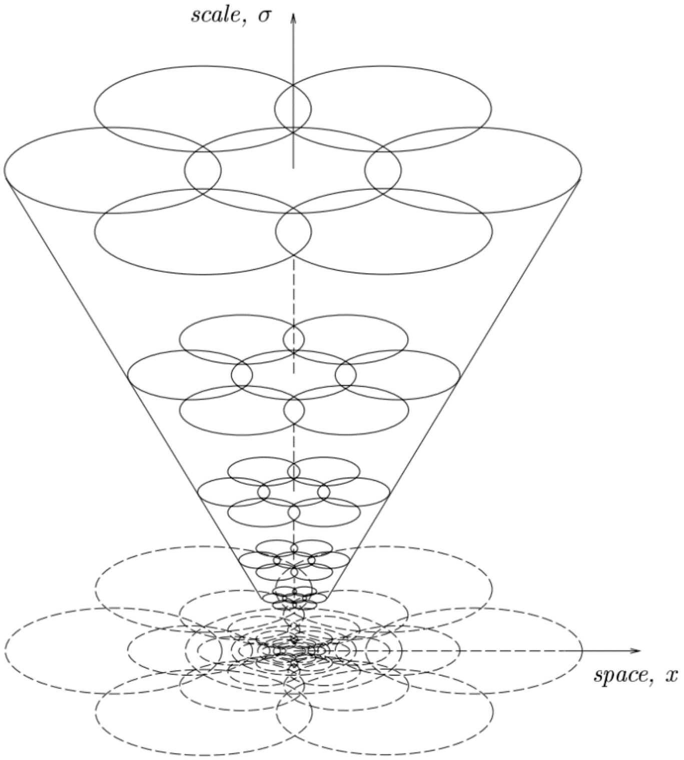


Figure 2-9 A schematic illustration of foveal system. It shows receptive fields at different level of the primate visual hierarchy. Note that the relative size of receptive fields is arbitrary. From (Lindeberg 1996).

2.5 Conclusion

In this chapter, we reviewed the literature that is closely related to the current work. The review starts from biological research. The intention is to provide a biological foundation to model early recurrence. Our understanding of the primate visual system is still an on-going process. This process leads to new theories that replace older ones. Further, theories motivate computational neuroscience with models that not only explain biological observations, but also influence research of computer vision. Biologically-inspired computer vision has therefore become popular in recent years.

Likewise, the main motivations for the current work from the biological side include the temporally asynchronized visual pathways in the primate visual system, and the lower-level cross-pathway feedback connectivity. These properties hint at a recurrent mechanism, which we called early recurrence that is fundamentally different from traditional attentive feedback mechanisms.

The second motivation is from the scale-space theory in computer vision, where the goal is to facilitate visual processing with scale-invariant visual representations. We noted that the scale-space methods focus on the hierarchical levels that are similar to where the early recurrent mechanism takes place. From this perspective, early recurrence may provide an additional solution to scale-invariant visual representations.

Chapter 3. A Computational Model of Early Vision with Recurrence

In this chapter, a computational model of early recurrent processing is proposed. As reviewed in the previous chapter, the main motivations from biology are the notion of lower-level cross-pathway feedback connections (Felleman and Van Essen 1991, Lienhart and Maydt 2002), and the fast-brain hypothesis (Bullier 2001). Our proposed model puts forth a modulating mechanism to improve early visual representation. The modulation takes place automatically as visual processing proceeds via the two visual pathways: the visual system uses results of computation from the dorsal pathway to modulate computation in the ventral pathway. The modulation takes form as a multiplicative inhibition process. After the modulation, early ventral information inconsistent with the dorsal representation is suppressed, leading to a refined version of the ventral representation for higher-level visual processing.

The computational foundations of the Selective Tuning (ST) model of visual attention explicitly permit the inclusion of early recurrence even though it was not part of the original formulation (Tsotsos 2011). There is no discussion of any automatic recurrent mechanisms in its previous works; however, they are also not precluded. The reason that automatic recurrent mechanisms such as those discussed here fit well within ST's attentional processing is that all of ST development is about reducing the search space inherent in the visual processing. This is precisely to what early recurrence contributes; the early and automatic recurrent influence described here helps in the reduction of the search for valid interpretations of early representations by reducing or eliminating the presence of spurious, and thus distracting in a search sense, responses.

In the first part of this chapter, a conceptual description of the model is introduced, followed by a discussion of how early recurrence is distinct from other feedback mechanisms in the literature. In the second part, a set of computational components to implement the proposed recurrent modulation is formalized. To fully explore the modulation, a feed-forward hierarchy is defined based on a simplified network of the two main visual pathways (reviewed in the previous chapter). Within this simplified network, two sets of recurrent connections are discussed: 1) recurrence from the dorsal area MT to the ventral layer of area V1, and 2) recurrence from the dorsal areas MT/MST to the ventral layer of area V2. Finally, using a synthetic example, we show how early recurrence improves the computation in the ventral pathway.

3.1 The Model

The model is defined using a simplified visual hierarchy as illustrated in Figure 3-1. The simplified hierarchy includes the two main visual pathways: the dorsal pathway and the ventral pathway. In this hierarchy, the dorsal pathway starts from the magnocellular layer of the LGN, continues via areas V1, MT, MST, and reaches areas 7a and FEF. The ventral pathway starts from the parvocellular layer of the LGN, reaches areas V1, V2, V4, and terminates at area IT. From the figure, we see that each visual area has an associated visual

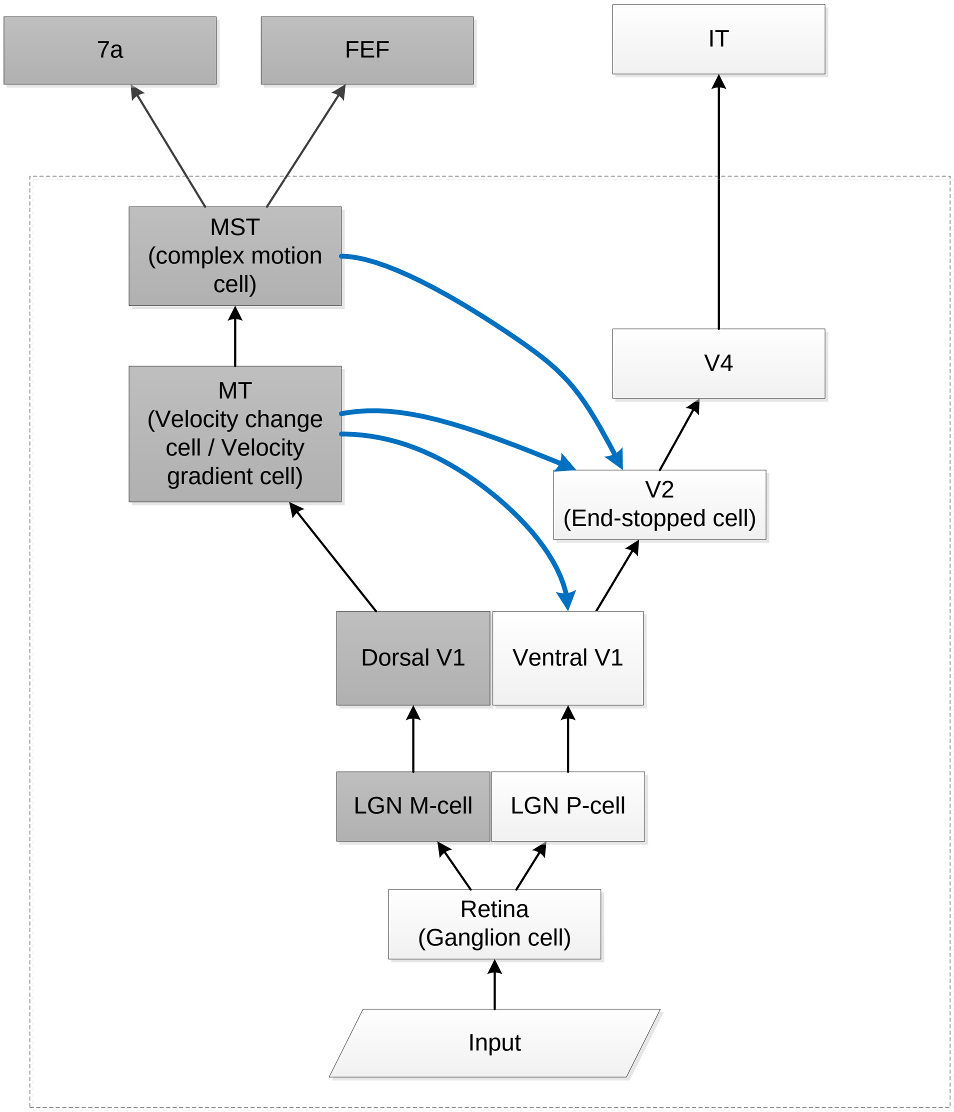


Figure 3-1 A simplified visual hierarchy. Areas in the dashed region are the deemed early visual areas, where early recurrence takes place (blue arrows). Black arrows are the feed-forward connections within the pathways. There are other types of feedback connections in the early visual hierarchy. However, they are not the focus of this study.

feature speciality (and this is also simplified and abstract for the purposes of this work). The hierarchy also includes a number of feed-forward and feedback connections. Further, each block in the figure conceptualizes a visual area. Blocks within the dashed region are the deemed early visual areas. Blocks out of the region are deemed as higher-level visual areas. They receive feed-forward activations from the early visual areas. The proposed early recurrent mechanism takes place within the dashed region. Visual areas outside this are not further considered in this work. Blue arrows illustrate the feedback connections. The biological support for the existence of these feedback connections can be found in the classic review of (Felleman and Van Essen 1991) and a recent update (Lienhart and Maydt 2002). Black arrows are the feed-forward connections within the pathways. It is worth noting that there are other types of feedback connections in the early visual hierarchy. However, they are not the focus of this study.

Although this view is much more simplified from the actual primate visual hierarchy, the network suffices to demonstrate the points of our model. The primary goal of this dissertation is within the computer vision domain; that is, we do not intend to develop a detailed biological vision theory. Instead, we put forth a computational model to describe a possible functional role for early recurrence, and ultimately to show its utility in improving computer vision algorithms. However, we will show in the next Chapter that the proposed model is capable of simulating neurophysiological experiments.

The proposed early recurrence leads to a cross-pathway feedback mechanism from early dorsal areas to early ventral areas. The conceptualized information flow of early recurrence is illustrated in Figure 3-2, where D-Li represents a dorsal area and V-Li represents a ventral area. Although many visual areas have bidirectional connections, and they may represent

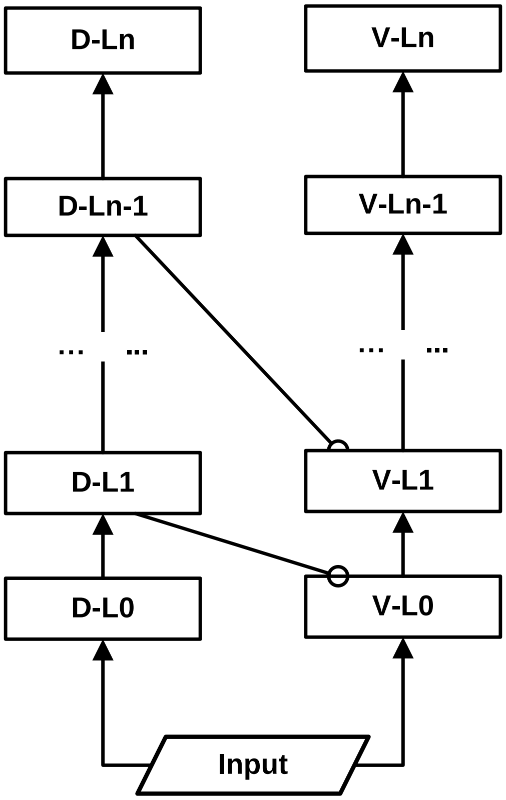


Figure 3-2 A conceptualization of early recurrence. Inputs are processed via two visual pathways. The dorsal pathway (D-L0 … D-Ln) has a faster processing speed. Thus its results could be sent back to the ventral area (V-L0 … V-Ln) before it receives feed-forward information from its lower levels. The figure particularly illustrates two such early recurrent possibilities, namely between D-L1 and V-L0, and between D-Ln-1 and V-L1.

pathways for a different kind of early recurrence, they are not considered in this thesis.

The timing of visual processing along the two pathways, as described in Chapter 2, has an asynchronising aspect. Visual areas within the dorsal pathway have shorter response latency to input stimuli compared with that of the ventral pathway. Given that there exist feedback connections between a dorsal area and a ventral area. Results of the dorsal area can reach the ventral area before it receives feed-forward information of the same stimuli from its lower-level visual area in the ventral pathway. Since the dorsal and the ventral pathways respond to different visual characteristics, feedback from the dorsal pathway provides the ventral pathway with information that is otherwise unavailable (from its own pathway). Figure 3-2 particularly illustrates two examples of such early recurrence, namely between D-L1 and V-L0, and between D-Ln-1 and V-L1.

Compared with existing models focusing on the interaction between the two visual pathways, the proposed model distinguishes itself in the hierarchical level of interaction, and also in the form of interaction. Let us focus on the first aspect, the hierarchical level of interaction. In its simplest form, assuming visual area on each level performs certain operations to extract visual information from its lower-level visual area. We use $\mathrm{d}_i()$ and $\mathrm{v}_j()$ to represent the operation of dorsal area i and ventral area j respectively. As illustrated in Figure 3-2, the chain reaction of the dorsal and the ventral pathway without early recurrence is then expressed as:

$$D_m = \mathrm{d}_m\left(\mathrm{d}_{m-1}\left(\ldots \mathrm{d}_j(\ldots \mathrm{d}_1(\mathrm{d}_0(I)))\right)\right), \tag{3-1}$$

$$V_n = \mathrm{v}_n\left(\mathrm{v}_{n-1}(\ldots \mathrm{v}_i(\ldots \mathrm{v}_1(\mathrm{v}_0(I))))\right), \tag{3-2}$$

where I is the input, and D_n and V_n are the outputs at the highest level of the two visual pathways respectively. Early recurrence happens at ventral layer 1. It leads to a refined visual representation. The chain reaction is altered, which can be described as:

$$\bar{V}_n = \mathrm{v}_n\left(\mathrm{v}_{n-1}\left(\dots\mathrm{v}_i\left(\dots\mathrm{v}_{1,\mathrm{d}_j}(\mathrm{v}_0(I))\right)\right)\right), \tag{3-3}$$

where $\mathrm{v}_{1,\mathrm{d}_j}$ represents that the ventral area v_1 is modulated by the dorsal area d_j. In this case, the top-level of the ventral pathway $\bar{V}_n$ is altered as a result of the chain reaction. Note that visual features extracted by the lower-level ventral areas are localized and relatively simple (e.g., edge, curvature). Therefore we propose that early recurrent modulation provides the ventral processing with a unique way to refine its representation from basic levels yet spatially accurate.

However, we must state that our early recurrent model has its limitations. The model itself describes an additional mechanism in the big picture of vision. We will discuss in the later chapters the similarities and the differences between the proposed model and prior arts. In addition, our understanding of the primate visual system is evolving; at present, we can only model certain aspects of early recurrence. However, we do believe the contribution of our model is significant in terms of helping to solve difficult computer vision challenges.

3.1.1 The Feed-forward Hierarchy

The simplified visual hierarchy (Figure 3-1) includes both the retino-cortical processing stage and the cortico-cortical processing stage. During the retino-cortical stage, retinal ganglion cells (RGC) project visual information to the primary visual cortex (area V1) via the lateral geniculate nucleus (LGN). Here, two types of RGC cells are modelled, the Parasol Ganglion Cell (PGC) and the Midget Ganglion Cell (MGC). The PGC has relatively larger dendritic tree and cell body than that of the PGC. Receptive fields (RFs) of the PGC and the MGC both have center-surround response patterns. The spatial extent of the PGC RF is in general larger than that of the MGC. The PGC and MGC have distinct contrast sensitivity (Kaplan and Shapley 1986). Temporally, the MGC responds to fast changing stimuli and the PGC responds to slowly changing variations.

In the LGN, two types of cells are modelled. They are the parvocellular (P-) cell and the magnocellular (M-) cell (Norton and Casagrande 1982). The P-cells receive most input from the MGCs. They have relatively smaller center-surround RFs, and respond to chromatic information. The M-cells receive most input from the PGCs. They have relatively larger center-surround RFs, and are blind to colour (Perry et al. 1984).

Figure 3-3 conceptually illustrates 2 cases of LGN center-surround RFs, on-centre and off-centre respectively. In each case, the LGN cell is the white disk on top of its input layer, and small disks represent RGCs. A small white disk represents on an RGC cell, and small grey disk represents an off cell.

From the LGN, visual processing enters the cortico-cortical stage. The first major visual area that receives input from the LGN is area V1 (Hendrickson et al. 1978, Lund 1988). The receptive fields of V1 neurons are larger than that of the LGN cells. LGN-V1 feed-forward connections are spatially aligned. V1 neurons are organized in hyper-columns (Ts'o et al. 1990), or the so-called ice cube structure (Hubel and Wiesel 1977). By this structure, V1 neurons exhibit an orderly progression spanning across different orientations.

Figure 3-4 conceptualizes the formation of a V1 spatial receptive field. Spatial orientation selectivity of the V1 neuron is determined by the spatial arrangement of its connected LGN cells. Each V1 neuron encodes one orientation. In this example, the V1 neuron encodes horizontal orientation.

From area V1, visual information projects to the two main visual pathways: the dorsal pathway and the ventral pathway (Mishkin et al. 1983, Casagrande and Xu 2004). The dorsal pathway starts from area V1, via area MT, MST, and terminates in the posterior parietal cortex. Functionally, this pathway is associated with motion, location and motor control. The ventral pathway starts from area V1, via area V2, V4, and terminates in the inferior temporal cortex (IT). This pathway is functionally associated with spatially-accurate object processing, such as edge, curvature, shape and object form.

In our model, each pathway is a multi-layered pyramid. Each layer is one of the aforementioned visual areas. It is further reasonable to assume each layer is organized as a grid of columnar computational units (neurons). Each unit receives input from units of other levels (via inter-cortical-region connections, feed-forward or feedback), or from units within the same level (intra-cortical-region connections, laterally). The two pathways and their visual areas are further discussed below.

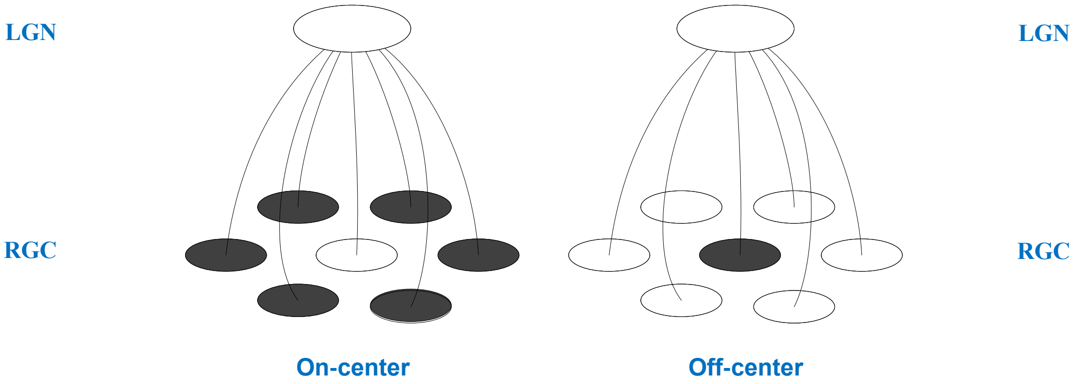


Figure 3-3 A semantic view of centre-surround receptive field of the LGN. Each LGN cell is the white disk on top of input layer cells (RGCs). Two types of LGN cells, on-centre and off-centre are modelled respectively.

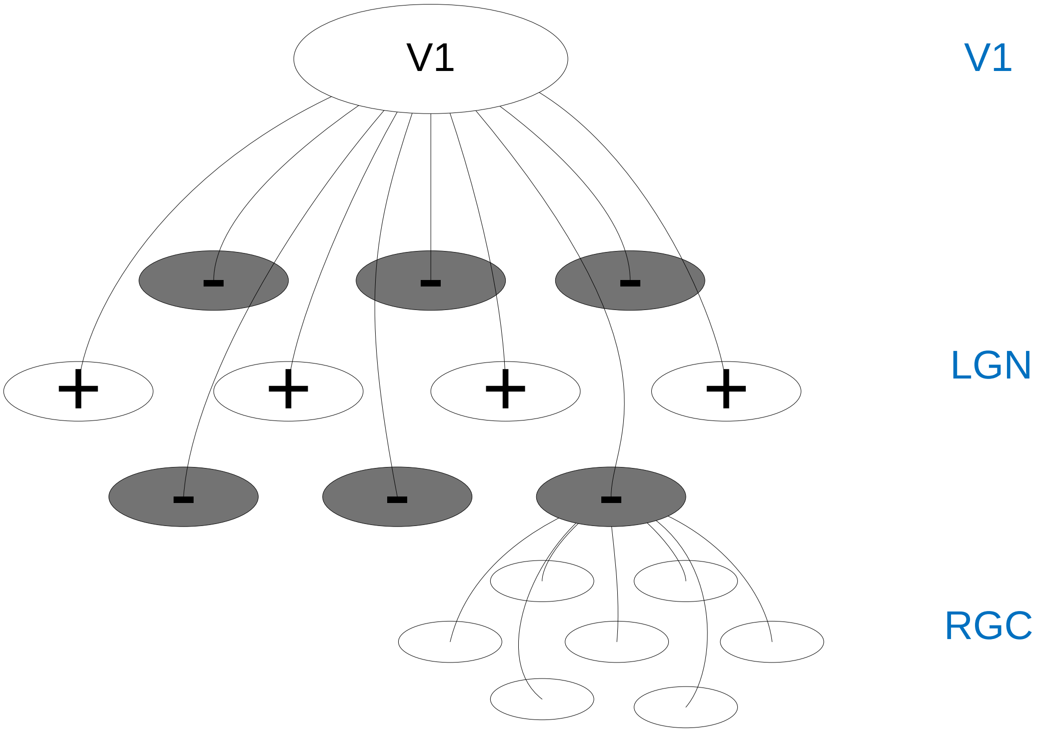


Figure 3-4 The formation of V1 spatial receptive field. Spatial orientation selectivity of each V1 neuron is determined by the spatial arrangement of LGN cells. In this example, the V1 neuron encodes horizontal orientation.

The Dorsal Pathway

The visual area MT receives most input from the dorsal layers of area V1 (Ungerleider and Desimone 1986). MT neurons are sensitive to spatiotemporal variations (Barberini et al. 2005). An MT neuron connects to a set of V1 neurons, and has a larger visual field. This property allows the MT neuron to integrate spatiotemporal information across larger spatial range. From an energy analysis perspective, this spatiotemporal response profile characterizes motion energy (Adelson and Bergen 1985). A direct result of the integration is the ability to achieve position-invariant motion perception, similar to the edge perception proposed in Neocognitron (Fukushima 1980, Fukushima 1988). Another result is to solve the aperture problem (Pack and Born 2001). In (Treue and Andersen 1996), the authors noted a spatiotemporal summation effect of MT neurons in macaques. In addition, they concluded a substantial portion of MT neurons respond to velocity changes (gradients), with each such neuron having a preferred spatiotemporal gradient orientation.

Based on these observations, the current work focuses on modelling two types of MT neurons (Figure 3-5). An MTvc neuron performs speed summation, which integrates V1 response of the same spatiotemporal orientation. That is, an MTvc cell has the spatiotemporal orientation selectivity identical to V1 neurons but with larger receptive field. An MTvg gradient-tuned neuron responds to spatiotemporal gradient changes. It receives input from a set of V1 neurons of different spatiotemporal orientations.

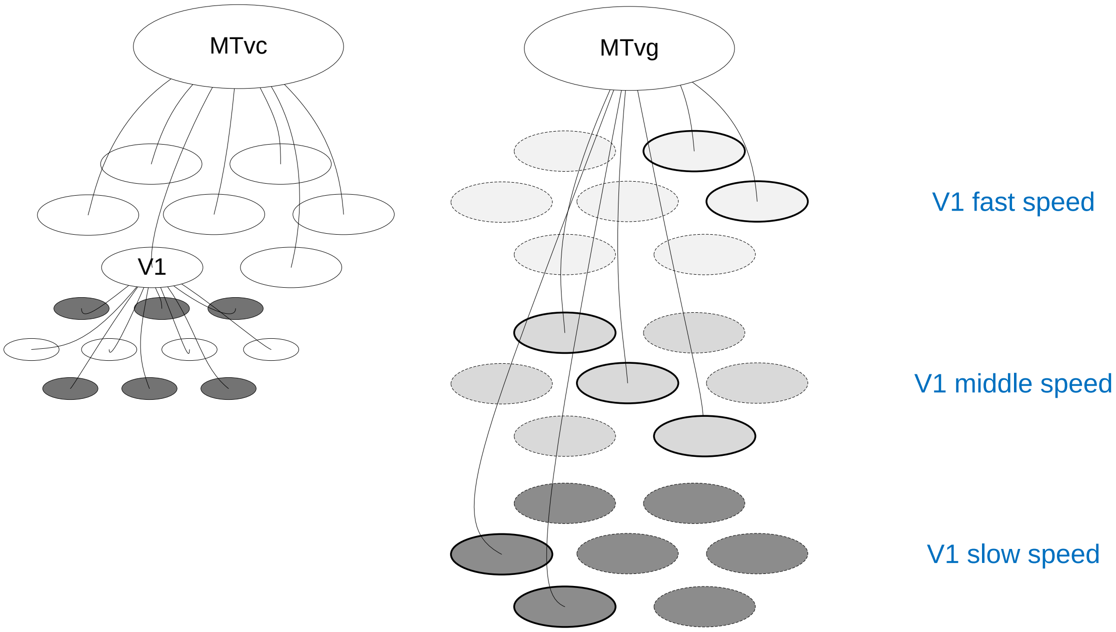


Figure 3-5 Two types of MT neurons are modelled. Left: flat speed summation MT neuron (MTvc). This neuron accumulates V1 responses of the same spatiotemporal orientation. Right: velocity gradient tuned MT neuron (MTvg). This neuron receives input from V1 of different spatiotemporal orientations. In this example, the MTvg integrates fast speed, middle speed and slow speed V1 neurons in horizontal direction.

Area MT connects to area MST (Maunsell and Van Essen 1983, Ungerleider and Desimone 1986). Each MST neuron receives input from a set of MT neurons, and thus has an even-larger receptive field. MST neurons respond to motion of different types, such as rotation, expansion, extraction, or combinations of these complex motion patterns. (Duffy and Wurtz 1991, Graziano et al. 1994, Duffy 1998). In principle, our model follows the motion model proposed in (Tsotsos et al. 2005).

The Ventral Pathway

Area V2 receives feed-forward projection from the ventral layers of area V1. It responds to end-stopped visual patterns (Dobbins et al. 1987). A V2 neuron can be modelled as a V1 simple cell inhibited by two displaced V1 complex cells at elongated ends (Figure 3-6). Depending on the orientations of displaced V1 complex cells, end-stopped cells respond to different curvatures. In (Rodriguez-Sanchez and Tsotsos 2012), four types of V2 neuron have been proposed. They respond to different local curvatures or corners. A Type-1 cell is the result of the sum of the responses of simple cells at the same location but with different orientations. A Type-2 cell has simple components that integrate the difference in response of two simple cells of different size at the same location. A Type-3 cell has complex components that result from the difference between a simple cell and two displaced complex cells. A Type-4 cell is the most generalized type; it adds a rotated component to a Type-3 cell such that it can distinguish between curvature directions. It is noted in (Rodriguez-Sanchez and Tsotsos 2011) that Type-1, Type-2 and Type-3 end-stopped cells are analogous to special cases of Type-4 cell. Thus, in the current work, we only implement Type-4 end-stopped cells.

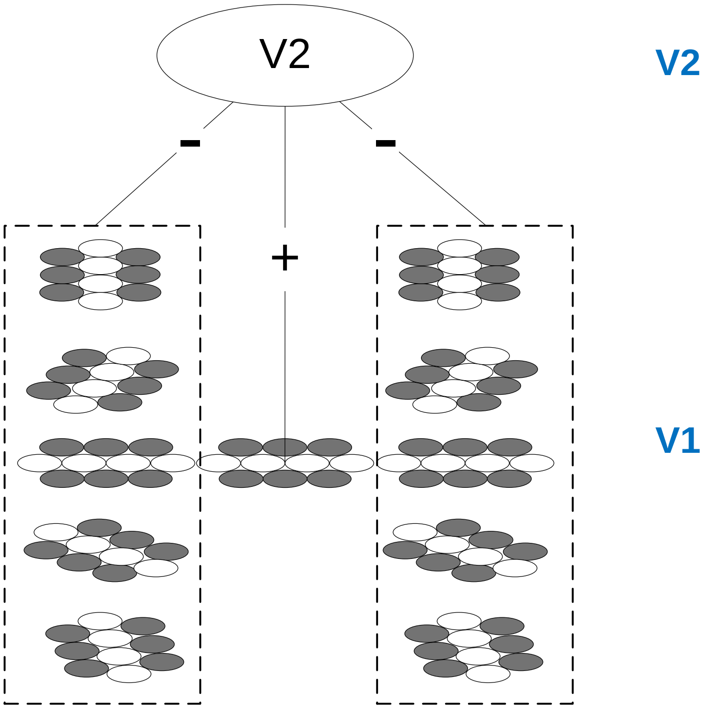


Figure 3-6 An example of a Type-4 V2 end-stopped neuron. End-stopped cell with complex components from the difference between a V1 simple cell and two displaced V1 complex cells (dashed enclosures) with additional rotated component.

3.1.2 Early Recurrent Modulation

The essence of our model lies in the recurrent connectivity and the modulation mechanism from a dorsal area to a ventral area within the early visual hierarchy. Specifically, early recurrence refers to the processing that applies the early dorsal representation to modulate early ventral processing.

Early recurrence is a mechanism of non-classical RF suppression. In Chapter 2, we have reviewed that classical RF refers to the region of visual space in which stimuli drive neural response. Non-classical RF, in contrast, is an extended surrounding region that can cause suppressive effect to neural response. Figure 3-7 schematically illustrates and compares the proposed suppression with non-classical RF inhibition mechanisms using object contour detection as an example. Self-inhibition (left figure) is an influential model of center-surround interactions. The idea has been supported by neurophysiological data and psychophysical experiments. Lower-level center-surround interaction mostly takes form as suppressive results (Nelson and Frost 1978). An orientation-selective cell will show reduced response when multiple stimuli of the same orientation present at the surrounding region (Schiller et al. 1976). In (Grigorescu et al. 2003), the authors demonstrated improved contour detection in cluttered.

Instead of the circular-shaped surrounding region, a butterfly-shaped surrounding region (Zeng et al. 2011) consists of two adaptive inhibitory end-regions and two non-adaptive inhibitory side-regions (Figure 3-7 middle figure). The butterfly-shaped inhibition zone includes two regions, a side region and an end region. The strength of the side region inhibition is calculated based on the local features in the side regions at a fine spatial scale, and the strength of the end region inhibition adaptively varies at both fine and coarse scales.

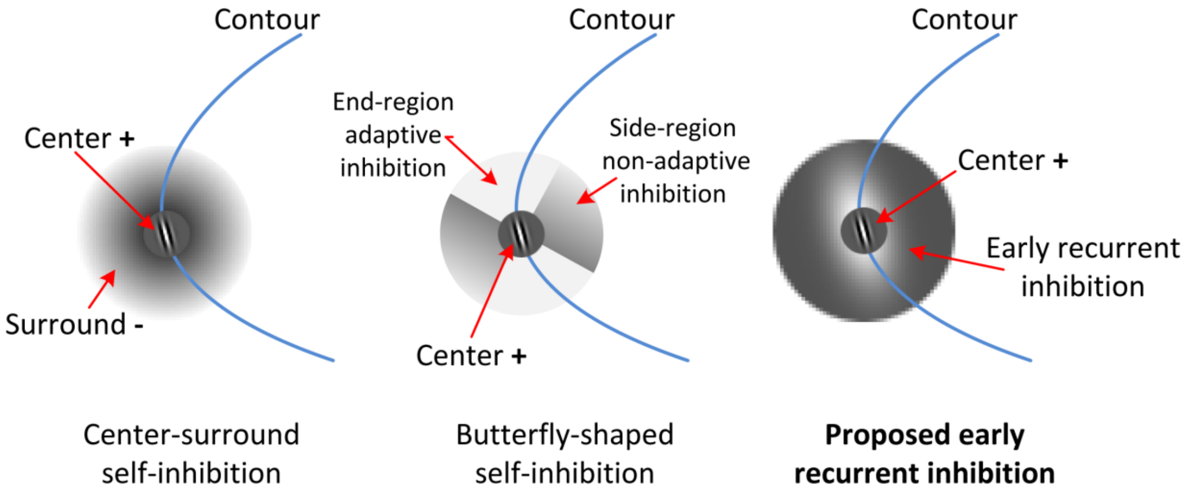


Figure 3-7 Schematic drawings of non-classical RF inhibition mechanisms. Left: self-inhibition (Grigorescu et al. 2003), Middle: the butterfly-shaped self-inhibition (Zeng et al. 2011), which consists of two adaptive inhibitory end-regions and two non-adaptive inhibitory side-regions. Right: the proposed recurrent inhibition.

Computationally, the end regions exert weaker inhibition where a contour is more likely to exist. The literature demonstrated that the object contours were extracted more effectively than (Grigorescu et al. 2003). However, the biological underpinning for such butterfly-shaped regions is not clear.

The most important aspect of the proposed early recurrent inhibition is that the strength of surrounding inhibition to the ventral neuron is caused by feedback from the dorsal pathway (Figure 3-7 right figure). The two pathways respond to different spatiotemporal visual characteristics. We can infer that early recurrence is capable of facilitating ventral computation with additional information that is not directly computable by the ventral pathway itself. This is in contrast to the aforementioned self- inhibition, where the inhibition representation is generated within the ventral pathway.

An important goal of this thesis work is to formalize the computation of early recurrence to improve computer vision applications. However, before deriving the computation, a discussion is needed to strengthen our proposed idea from biological perspectives. In particular, the work has three hypotheses. The first hypothesis concerns the conditions required to the visual network to accommodate and exert early recurrence. This is the precondition of early recurrence. The second hypothesis concerns the nature of operation of early recurrence. This and the first hypothesis together will lead us to formalize the computational components of early recurrence. The third hypothesis concerns the functional role of early recurrence. This hypothesis inspires us to apply early recurrence to improve computer vision applications. In all, these three hypotheses distinguish early recurrence from other known recurrent mechanisms. In what follows, we will describe these hypotheses and related biological research to support them. Specifically, since Hypothesis 1 concerns the biological facts, we provide evidence for it by reviewing important research from the literature. In Chapter 4, we apply the model to simulate two biological experiments. Observations from the simulations will also be used as evidence to validate Hypothesis 2 and Hypothesis 3.

Hypothesis 1: Requirements

In the primate visual system, two visual areas, with one in the dorsal pathway and the other in the ventral pathway, must satisfy the following two requirements, without which early recurrence cannot take place. The first requirement is that there *exists a feedback connection* between a dorsal area (source) and a ventral area (target). The second requirement is that *timing of processing* allows the feedback signal from the source to reach the target before the target receives the feed-forward signal, which is via the ventral pathway.

The two requirements are illustrated in Figure 3-8. In this example, D_i refers to the dorsal region at level i. V_j refers to the ventral region at level j. Note that in this case, level j is at a lower level than level i in the visual hierarchy. ffd denotes the feed-forward connectivity in the dorsal pathway and ffv denotes the feed-forward connectivity in the ventral pathway, and fb denotes the feedback connectivity. The timing condition manifests itself as the following formula:

$$T_{ffd} + T_{D_i} + T_{fb} \leq T_{ffv}, \tag{3-4}$$

where T_{ffd} refers to time required for signals of visual input to travel to the dorsal region, T_{D_i} refers to time required for the dorsal region D_i to respond to the input, T_{fb} refers to time required for the dorsal responses to travel back to the ventral region V_j, and T_{ffv} refers to time required for signals of visual input to travel to the ventral region.

We found evidence to support the first requirement from brain anatomy studies using combined injection techniques. Van Essen and colleagues concluded the existence of not one but numerous cross-pathway recurrent connections in the primate visual system (Maunsell and Van Essen 1983, Van Essen and Maunsell 1983). One possible dorsal region is area MT, which receives input from area V1 and has feed-forward connections to higher-level visual areas, such as area MST. Being at the center of the dorsal pathway, studies revealed that area MT has feedback connections to many visual areas in the ventral pathway, such as areas V1, V2, and V4. Further, feedback effects have been observed in both non-human primates (Salin and Bullier 1995, Hupé et al. 2001) and humans (Pascual-Leone and Walsh 2001, Silvanto et al. 2005).

Feedback connections from area MT to lower-level visual areas show diversified patterns. This indicates that feedback signals might modulate the visual hierarchy in different manners. If feedback is from a dorsal area and feedback axons terminate at a ventral area, then the first condition of our hypothesis is satisfied. For example, the feed-forward connection from area V1 to area MT is mostly from layer 4B of area V1. However, feedback from area MT reaches layer 6 of area V1 (Maunsell and Van Essen 1983). This indicates an asymmetric feed-forward-feedback loop, which makes it possible that area MT feedback has a modulatory effect not only on the dorsal layers of area V1 (4Cβ), but also the ventral layers of area V1 (4Cα). Similar asymmetric loops have been found between area MT and area V2: feed-forward starts from layer 2 and layer 3 of area V2, and feedback starts from area MT and terminates at almost all layers of area V2 (Maunsell and Van Essen 1983).

Several studies examining the timing of visual processing provide evidence that the second requirement is also satisfied. Since mid-1990s, multi-sites cell recording techniques

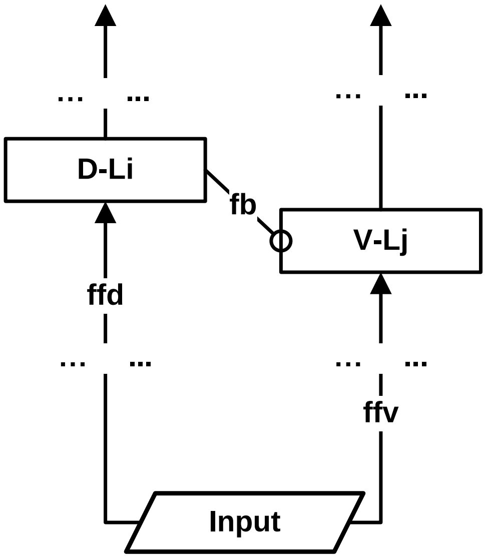


Figure 3-8 Early recurrence requires that: 1) there exists feedback connections: ffd and ffv are feed-forward connections, and fb is the feedback connection; and 2) the accumulated time of feed-forward projection (T_{ffd}), of neuron responding (T_{fb}), and of feedback projection along the dorsal pathway (T_{fb}) is no more than the time needed for feed-forward traversal along the ventral pathway (T_{ffv}).

on rats, cats and primates have provided us with significant observations of temporal processing order. Bullier and his colleagues (Henry et al. 1991, Salin and Bullier 1995, Nowak et al. 1997, Angelucci and Bullier 2003) conducted several neurophysiological probes and confirmed the temporal order of lower-level cortical processing at a number of visual areas. During the same period of time, Schmolesky and colleagues (Schmolesky et al. 1998) reached similar conclusions using macaque monkeys. With a unified test protocol over multiple visual areas, they concluded that the dorsal pathway and the ventral pathway respond to visual stimuli with very different time courses. In general, visual areas in the dorsal pathway respond to visual input much earlier than those in the ventral pathway. Specifically, the magnocellular cells of the LGN become active 15-20 milliseconds earlier than that of the parvocellular cells in macaques (Maunsell et al. 1999). The authors pointed out that this temporal separation leads to a diversified representation in layers 4Cα and 4Cβ of area V1. Thus, it is suggested that a functional separation in the LGN and area V1 appears to have more profound impact on the dorsal and ventral pathway beyond area V1.

More recently, consistent results have been observed on humans using magnetic and imaging techniques. These non-invasive techniques are essentially harmless. By combining the two, one can collect accurate results, spatiotemporally. Based on several studies, it was concluded in (Barnikol et al. 2006) that the critical time for the first sweep of object processing in the visual hierarchy is around 50 -150 milliseconds after stimuli onset. The study also noted that there are strong recursively deployed cortical arrangements. We infer that this short-delayed recursion cannot come from higher-level cortical areas, but caused by lower-level recurrence. Based on the literature, a list of visual areas that are capable of accommodating early recurrence is derived (Table 3-1).

Table 3-1 Visual areas that could allow early recurrence to take place. Time measurements in brackets indicate response delays after stimuli onset. Numbers are concluded from the literature as listed below. "–" indicates that the literature does not support recurrent connectivity between the regions.

From To	MT (45msec)	MST (45msec)	7a (90msec)
V1 (65msec)	*Likely*	*Likely*	Not Likely
V2 (85msec)	*Likely*	*Likely*	Not Likely
V4 (100 - 150msec)	–	*Likely*	*Likely*
IT (100 - 400msec)	–	–	*Likely*

*** Literature provides support and the indicated species:**

Area V1: (Maunsell and Gibson 1992) monkey, (Nowak et al. 1995) monkey

Area V2: (Nowak et al. 1995) monkey

Area V4: (Chelazzi et al. 2001) monkey

Area IT: (Goebel et al. 1998) monkey

Area MT: (Pascual-Leone and Walsh 2001) human, (Schmolesky et al. 1998) monkey

Area MST: (Kawano et al. 1994) monkey, (Schmolesky et al. 1998) monkey

Area 7a: (Bushnell et al. 1981) monkey

Hypothesis 2: Recurrent Operation

The Hypothesis 2 is that early recurrence modulates a ventral representation in a *surround suppression* fashion. In the context of visual search, it has been shown that top-down recurrent processing causes surround suppression (Tsotsos 1990, Boehler et al. 2009). Such top-down attentive suppression is a key component of Selective Tuning (ST) (Tsotsos et al. 1995). The authors proposed a mathematical representation, theta-winner-take-all (θ-WTA). The operation suppresses irrelevant features within a receptive field. θ-WTA proceeds in successive layers in a top-down fashion, leading to an inhibition beam. At the lowest level of this beam (input layer), pixels attracting the attention are localized.

Although ST is a model of visual attention, the surround suppression computation is consistent with the current work, as is its motivation. However, the functional role of early recurrence in our work does not concern attentional processing per se (that is, top-down volitional attention), attended visual features or attended spatial locations. Rather it plays a simpler role to improve the quality of ventral processing. Early recurrence has no top-down component and occurs automatically along the feed-forward projection rather than deliberately. Therefore, as will be shown in the formalization section, the θ-WTA has been much simplified to exclude attention-related parameters.

We propose that early recurrence operates non-linearly and multiplicatively. The non-linearity is two-fold. The first is the non-linear response pattern of the dorsal neurons: given a linear input (e.g., contrast), a neuron's response saturates at a certain threshold, leading to non-linear output strength. This responding profile has been observed in the early dorsal areas and has been modelled as a sigmoid function or a softmax function (Snowden et al. 1991, Simoncelli and Heeger 1998, Osborne et al. 2004).

Another factor causing non-linearity is due to the feedback connectivity between the dorsal neurons and the ventral neurons. As noted in (Angelucci and Bullier 2003), a ventral neuron has connection with several dorsal neurons, and vice versa. If feedback influences the ventral computation linearly, one should observe ventral neural activity changes regardless to feed-forward activations. However experiments suggested the opposite results (Girard and Bullier 1989, Girard et al. 1992): only feed-forward activation has such a linear response pattern. When connections from area V1 to area V2 are blocked, V2 neurons do not respond at all, despite the fact that they receive feedback activations from the dorsal area MT.

It should be clarified that the actual feedback representation and the actual modulation is complicated. The current work does not intend to characterize its full scope. Instead, we are interested in exploring how to behaviourally formalize the mechanism, such that one can use it to improve computer vision algorithms in practice.

We propose that one way to define the recurrent representation is via a weighted-sum operation. The weights can base on multiple criteria, such as connectivity and neural saturation. The recurrent representation is defined as the point-by-point measure of degree of influence of a dorsal neuron on a ventral neuron. Numerically, a large value indicates a strong dorsal response, while a small value indicating a weak dorsal response. A value A_i in the recurrent representation can be defined as:

$$A_i = \sum_{j=1}^{n} \omega_j \alpha_{ij}, \tag{3-5}$$

where $j = 1, \dots, n$ represents a criterion, $i = 1,2,3, \dots, m$ indexes neurons, α_{ij} denotes input strength of neuron i at criterion j, and ω_j denotes weight.

Given a recurrent representation, it is proposed that the ventral neuron is modulated via multiplication, which is defined as:

$$R_i = |\ A_i * I_i\ |, \tag{3-6}$$

where I_i denotes the input strength (ventral neuron), R_i represents the modulated result, $*$ represents the modulation, and | | is a rectification function to bound output strength within certain range. In a simple way, the recurrent representation defines inhibition strength but in a reverse manner. A large value in R_i indicates weak inhibition strength, and a small value indicates strong inhibition strength.

In the formalization section, different possible types of recurrent operations are considered and compared. A quantitative analysis shows that multiplicative inhibition fit best to biological data.

Hypothesis 3: Functional Role

The Hypothesis 3 is that the functional role of early recurrence is to apply dorsal representation to *refine* ventral representations. Earlier, a connection to scale-space theory was briefly described but not further pursued. That connection has relevance here. In essence, the dorsal and ventral pathways perform computations at different spatial scales (as well as temporal) as commonly considered in scale-space theories in computer vision. The effect applies beyond edges, however, and is true for any visual features that may exist over different spatial (or temporal) scales. In this sense, we may assert that perhaps biological vision is specifically designed to embody a form of scale-space analysis and our work is an instance of this.

A Contextual Modulation View of Early Recurrence

Before presenting computational details, it is worth mentioning the significance of early recurrence to the big picture of vision. At first glance, early recurrence is a shortcut of the visual hierarchy and a hardwired mechanism. The routing pattern between the two pathways seems not to conform to the Hubel and Wiesel's hierarchy. Early recurrence provides the ventral pathway a form of local context. In later chapters, we will formalize this local context in two forms: a spatiotemporal context, i.e., motion information, and a coarse-scale spatial context. Synthetic and real images will be used to investigate the influences of these contexts on visual processing in a number of neurobiological experiments and computer vision applications.

The proposed early recurrent mechanism is very different from global contextual modulation models (Bar 2004, Oliva and Torralba 2007). In these works, the authors described a global context representation for object recognition. The common characteristics between Bar's model and the current work is that in both models, the context representation is generated in the dorsal pathway. The proposed early recurrence is different from it from at least two aspects. The first aspect is the form of context representation. In Bar's theory, context is a description of spatial stimuli arrangements. It is built on a global view, which is not likely to happen for the early dorsal areas. The second aspect is the context facilitation operation. In Bar's theory, context directly biases the object selection process, which takes place in the Inferotemporal complex (IT). The essence of the global context is a guess of object categories (Bar 2004, Kveraga et al. 2007) or a probabilistic deployment of how likely the object may be found in the ventral presentation given the global description (Torralba 2003, Torralba et al. 2003, Oliva and Torralba 2007). The global contextual modulation enables the visual system to prune unlikely object candidates. Implicitly, this mechanism relies on past experience. Performance will therefore degrade given unfamiliar object or scene settings.

In our work, we consider an operation at much lowered visual areas. Our work concerns what local spatiotemporal scale differences may be productively combined to improve the representations upon which higher order computations, such as those in area IT, operate.

Examples of Early Recurrence

Our first example considers early recurrence between area MT and area V1/V2. Area MT is concerned with integration of velocity and velocity gradient from visual area V1 (Ungerleider and Desimone 1986, Treue and Andersen 1996, Born and Bradley 2005, Rust et al. 2006). Early recurrent connections have been located between area MT and area V1 (Figure 3-9), and between area MT and area V2 (Figure 3-10). From Table 3-1, the timing of visual processing allows feedback from area MT to reach area V1 and area V2 before they receive feed-forward signals from the ventral connections. Therefore, we propose that early recurrence between area MT and these early visual areas exist. Specifically, for area V1, the modulation refines the V1 representation of spatial orientation, and for area V2, the modulation improves the V2 end-stopped cell responses.

Our second example is early recurrence between area MST and area V2. Area MST receives most feed-forward inputs from area MT. Neurons in area MST have large receptive fields and thus respond to large-scale spatiotemporal variations, motion patterns, or optical flows (Duffy and Wurtz 1991, Duffy and Wurtz 1997, Duffy 1998, Smith et al. 2006). More specifically, motion patterns extracted by area MST include translation, rotation, extraction, and contraction (Graziano et al. 1994, Watanabe 1998). However, as a higher-order visual area, area MST has a coarse spatial representation, which means the representation lacks spatial accuracy: although responding actively to motion, area MST is unable to localize the moving stimuli.

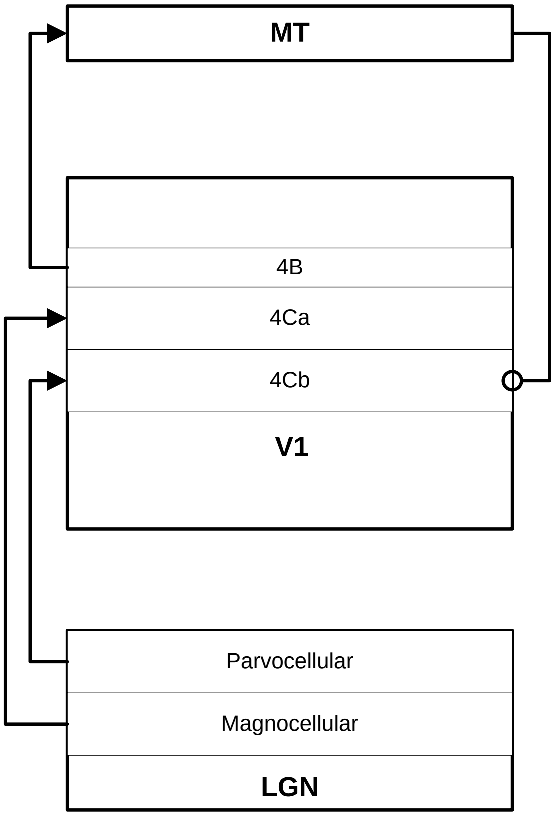


Figure 3-9 Recurrent connections from MT to area V1 (layer 4Cβ). During the dorsal processing, feed-forward information reaches area MT via area V1 (layer 4Cα-4B) approximately 45 milliseconds after stimulus onset. The temporal difference allows results of MT to modulate area V1 (layer 4Cβ) approximately 65 milliseconds after stimulus onset.

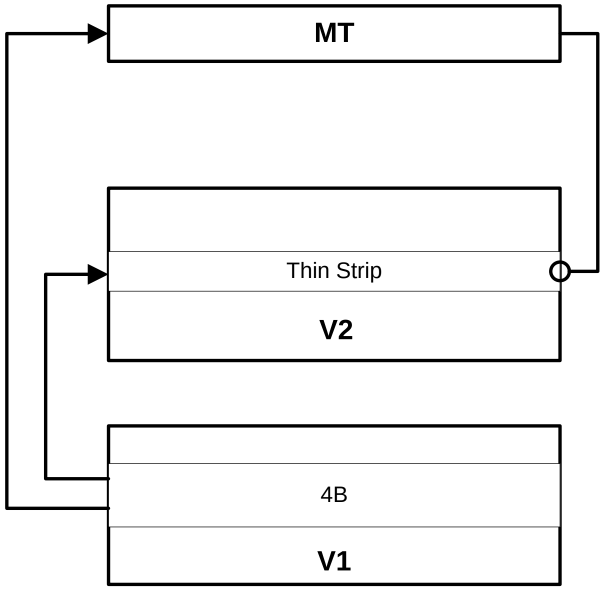


Figure 3-10 Recurrent connections from area MT to area V2 (thin stripe). During the dorsal processing, feed-forward information reaches area MT via V1 (layer 4Cα-4B) approximately 45 milliseconds after stimulus onset. The timing allows result of area MT to modulate area V2 (thin stripe) approximately 85 milliseconds after stimulus onset.

As reviewed in the previous chapter, the literature supports that area MST has recurrent connections to area V2 (Figure 3-11) (Duffy and Wurtz 1991, Felleman and Van Essen 1991, Graziano et al. 1994) and that the timing (Table 3-1) permits recurrence from area MST reaching area V2 before area V2 receiving feed-forward activation from area V1 (Schmolesky et al. 1998). Therefore, we proposed that there exist early recurrent mechanisms between area MST and area V2, and that feedback from area MST improves curvature representation computed by the end-stopped cells found in area V2 (Levitt et al. 1994, Felleman et al. 2015). The modulation turns curvatures inconsistent with area MST's representation inhibited. Modulation improves the signal-to-noise ratio of curvature representation, where the signal refers to actual curvature stimuli and noise refers to distracting stimuli. It is thus the modulated curvature representation that projects into higher-level ventral areas for object processing.

Early Recurrence Patterns

It is proposed that the early recurrent operation depends on the axon terminating patterns between the dorsal neurons and targeted ventral neurons. As shown in Figure 3-12, the current work sketches three modulation patterns. To clearly present the patterns, we use MT-V1 recurrence as an example:

Pattern-1: in its simplest manner, early recurrence has a spatially anisotropic suppression pattern. Each neuron of area V1 is modulated by recurrent signals from a set of neurons from area MT of the same spatial orientation followed by a sigmoid non-linear operation (Figure 3-12 top figure).

Pattern-2: each neuron of area V1 is modulated by recurrent signals from a multitude of neurons of area MT of different spatial orientations. Each neuron of area MT contributes to the modulation with equal strength (Figure 3-12 middle figure). In this case, the recurrence is spatially isotropic. The recurrent representation reflects the motion energy calculated by the dorsal pathway. The recurrence inhibits response of area V1 if recurrent signals from area MT indicate no motion in the neighbourhood of area V1. Alternatively, neural activity of area V1 remains unaffected if recurrence indicates a strong local motion pattern (without regards to any specific direction or direction gradients).

Pattern-3: in a more generalized manner, a neuron of area V1 can be modulated by recurrent signals from a multitude of neurons of area MT of different spatial orientations. Unlike Pattern-2, each neuron of area MT contributes to the modulation with varying strength depending on the connectivity (Figure 3-12 bottom figure). It is easy to see that Pattern-3 is the generalized form, where Pattern-1 and Pattern-2 are two extreme cases. In reality, each V1 neuron has a preferred tuning profile. Maximum suppression to the neuron may be achieved if the recurrent signal is at its preferred orientation. Conversely, a V1 neuron is least suppressed (or not affected at all) if the recurrent representation is not at its preferred orientation. Numerically, the modulation at a V1 neuron can be modelled as weighted sum over all connected MT neurons.

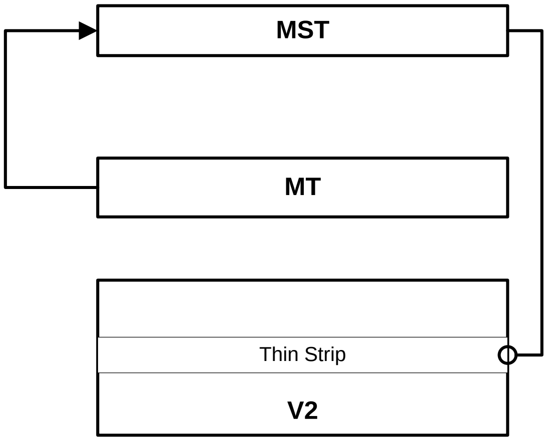


Figure 3-11 Early recurrent connection from area MST to area V2 (thin stripe). During the dorsal processing, feed-forward information reaches area MST via area V1 (layer 4Cα-4B) and area MT approximately 45 milliseconds after stimuli onset. Temporal properties allow results of area MT feed back into area V2 (thin stripe) to modulate ventral processing that starts approximately 85 milliseconds after stimuli onset.

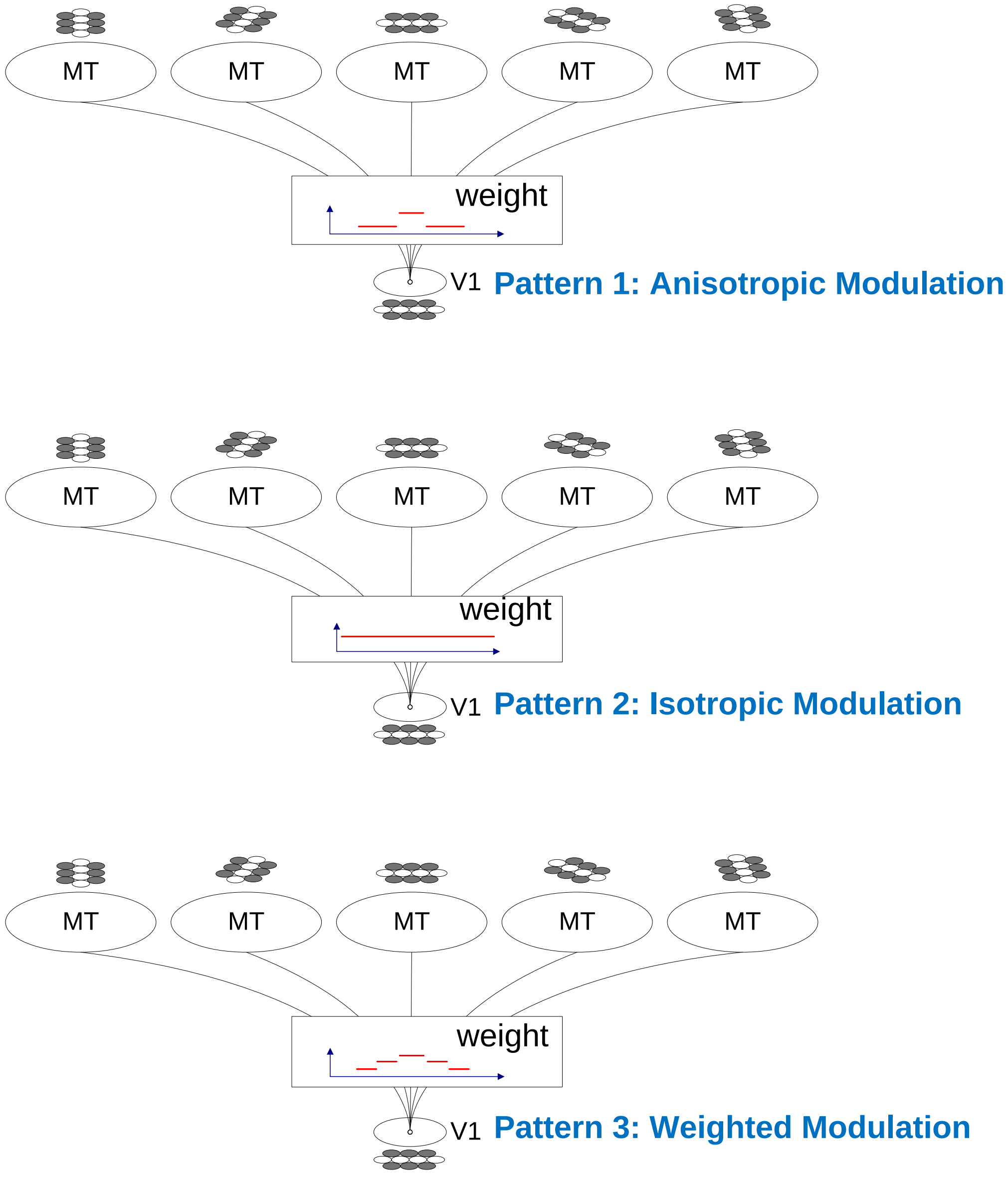

MT
MT
MT
MT
MT
weight
V1
Pattern 1: Anisotropic Modulation
MT
MT
MT
MT
MT
weight
V1
Pattern 2: Isotropic Modulation
MT
MT
MT
MT
MT
weight
V1
Pattern 3: Weighted Modulation

Figure 3-12 Three axon-terminating patterns lead to three early recurrent modulations paradigms. Coordinate with red bars represent weighting strength.

3.2 Formalization

This section formalizes the computational components for the proposed early recurrent modulation. We first define the receptive field of each visual area in the simplified hierarchy using an image-filtering approach. To best associate a filter representation with the formalized receptive field, the filter is implemented and parameterized informed by the relevant neurophysiology. In our simplified visual hierarchy, the two feed-forward pathways start from the retina, via the LGN, and forward into a number of cortical regions. These cortical regions include area V1, V2, MT, and MST. With an example using synthetic octagon stimuli, we will show how visual features are extracted and interpreted along the two visual pathways. Note that we will not show recurrence results in this section, as they will be discussed in detail in the following chapters.

On the basis of this filter-based visual hierarchy, the computation of early recurrence is formalized. Early recurrent representation is in general a weighted summation over responses of a set of dorsal neurons and its application takes the form of multiplicative inhibition. In what follows, computational relationships between neurons in the dorsal pathway (source neurons) and neurons in the ventral pathway (target neurons) are formalized in detail. Numerically, we derive anisotropic and isotropic modulation paradigms, which correspond to Pattern-1 and Pattern-2 as discussed in the previous section, respectively.

The visual pathways considered in the current network are illustrated in Figure 3-1. To facilitate our discussion, the word “modelled” is used hereafter as an adjective prior to the word “cell” or “neuron” to indicate that the entity is a computational component proposed in the current work, whereas a cell or neuron without the word “modelled” implies a biological entity.

As reviewed above, neurons of higher-level visual areas have larger receptive fields: they "see" larger visual areas than lower-level neurons. From a connectionist perspective, this is achieved by network convergence, where a higher-level neuron receives inputs from multiple lower-level neurons. To make our computation follow this property, we use relative receptive field size as a factor to define filter kernel size. The relative receptive field size is realized by relative filter kernel size. A size ratio (SR) is thus defined to describe the relative size of the receptive field with regards to the lowest level receptive field.

Table 3-2 provides an informative comparison of RFs in different visual areas. Sizes reported in the literature are on different animals. It is therefore difficult to make a fair comparison across different layers of areas, except for areas V1, V2 and MT. Nevertheless, we can infer from the table that as information progresses to higher level visual areas, size of RF increases. The finest spatial accuracy is at Midget Ganglion Cells (MGCs) in the retina. MGCs connect to P-cells in the LGN, and the relative receptive field size is approximately 2.9°. At area V2, receptive field increases to 5.4°, and size ratio (SR) is 18, which indicates that the dimension of the filter representing a V2 neuron should be about 18 times larger than that of the MGC. From the table, we can see that at the level of area MT, each neuron responds to a spatial region of approximately 15.3°, and has a SR of approximately 51 times that of MGC.

Table 3-2 Comparison of receptive field size

Cell/Neuron	**Average receptive field center diameter**	**Size Ratio (SR)**
RGC (MGC)	0.3° (Peichl and Wassle 1979) cat	1.0
RGC (PGC)	0.83° (Peichl and Wassle 1979) cat	2.7
LGN (P-cell)	0.87° (Xu et al. 2001) owl monkey	2.9
LGN (M-cell)	0.92° (Xu et al. 2001) owl monkey	3.1
V1	0.5° - 1.3° (Angelucci et al. 2002) macaque	1.5 - 4.5
V2	5.4° (Angelucci et al. 2002) macaque	18
MT	15.3° (Angelucci et al. 2002) macaque	51

3.2.1 Formalization of the Feed-forward Pathways

We model the feed-forward processing in the simplified visual hierarchy as an image filtering system involving a cascade of image filters. Each visual area is modelled by a set of image filters that characterize it receptive field in the classical way.

3.2.1.1 The Retinal Ganglion Cell (RGC)

Two types of retinal ganglion cells are modelled. The Parasol Ganglion Cells (PGCs) and the Midget Ganglion Cells (MGCs) have center-surround response properties with different spatiotemporal frequencies. The spatial response profile of the modelled PGCs and the modelled MGCs is defined by a two-dimensional Difference-of-Gaussians (DoGs) function derived from (Rodieck 1965, Davson 2012):

$$f_{RGC}^{spatial}(x,y) = \frac{1}{2\pi\sigma_c^2} e^{-\frac{x^2+y^2}{2\sigma_c^2}} - \frac{1}{2\pi\sigma_s^2} e^{-\frac{x^2+y^2}{2\sigma_s^2}}, \tag{3-7}$$

The spatial frequency of RGC varies by changing the value of σ_{RGCs} (Figure 3-13 B). As suggested in (Marr and Hildreth 1980), we use $\sigma_{RGCc} = 1.6 * \sigma_{RGCs}$, which approximates one-octave bandwidth (Figure 3-13 C).

Receptive fields of MGCs (Figure 3-13 plot A) and PGCs (Figure 3-13 plot B) have similarly shaped spatial profiles but different spatial frequency tuning properties. In general, the PGCs have relatively large RF size, which respond to low-spatial-frequency input. In the current work, we use $\sigma_{RGCs} = 0.8$ and 1.6 for MGCs, and use $\sigma_{RGCs} = 1.6$ and 3.2 for PGCs. Plot C and plot D compare cell response profile in spatial domain and frequency domain respectively. As σ increases, the spatial region of the receptive field extends, and its frequency selection becomes lower.

The RGCs respond after receiving a preferred signal. The temporal delay is short. The temporal characteristic of RGCs is modelled using a one-dimensional low-pass filter:

$$f_{RGC}^{temporal}(t) = \frac{\mathrm{T}(t)}{2\pi\sigma_{RGCt}^{2}} e^{-\frac{t^{2}}{2\sigma_{RGCt}^{2}}}, \tag{3-8}$$

where $\mathrm{T}(t) = 1$ iff $t > 0$; this forces $f_{V1}^{temporal}(t) = 0$ when $t \leq 0$.

Therefore, given an input stimuli $I(x, y, t)$, the response of RGCs is represented as:

$$R_{RGC}^{spatiotemporal} = \Theta\left[I(x, y, t) \circledast f_{RGC}^{spatial}(x, y) \circledast f_{RGC}^{temporal}(t)\right], \tag{3-9}$$

where $\circledast$ denotes the convolving operation, and $\Theta[m]$ denotes a sigmoid function that rectifies the responses in a non-linear manner using a Hyperbolic tangent function.

3.2.1.2 The Lateral Geniculate Nucleus (LGN)

The LGN receives feed-forward input from the RGCs. Two types of LGN cell, the magnocellular cell (M-cell) and the parvocellular cell (P-cell) are formalized. The third type, the Koniocellular cell, is ignored in this work. This is because the property of K-cell is somewhere between the M-cell and the P-cell. Thus, by formalizing the M-cell and the P-cell, our computational representation is sufficient to express the main idea. It might be the case that K-cells add robustness due to the additional scales of space and time considered but this examination is left for future work. In our work, modelled M-cells receive input from modelled PGCs, and modelled P-cells receive from modelled MGCs.

Similar to the RGCs, the LGN cells have center-surround receptive fields. Unlike the RGCs, receptive fields of the LGN cells are relatively large. The spatial response patterns (Figure 3-14 plot A and plot B) of the modelled M-cell and modelled P-cell are determined as a two-dimensional Difference-of-Gaussian function (Rodieck 1965, Einevoll and Plesser 2012) as:

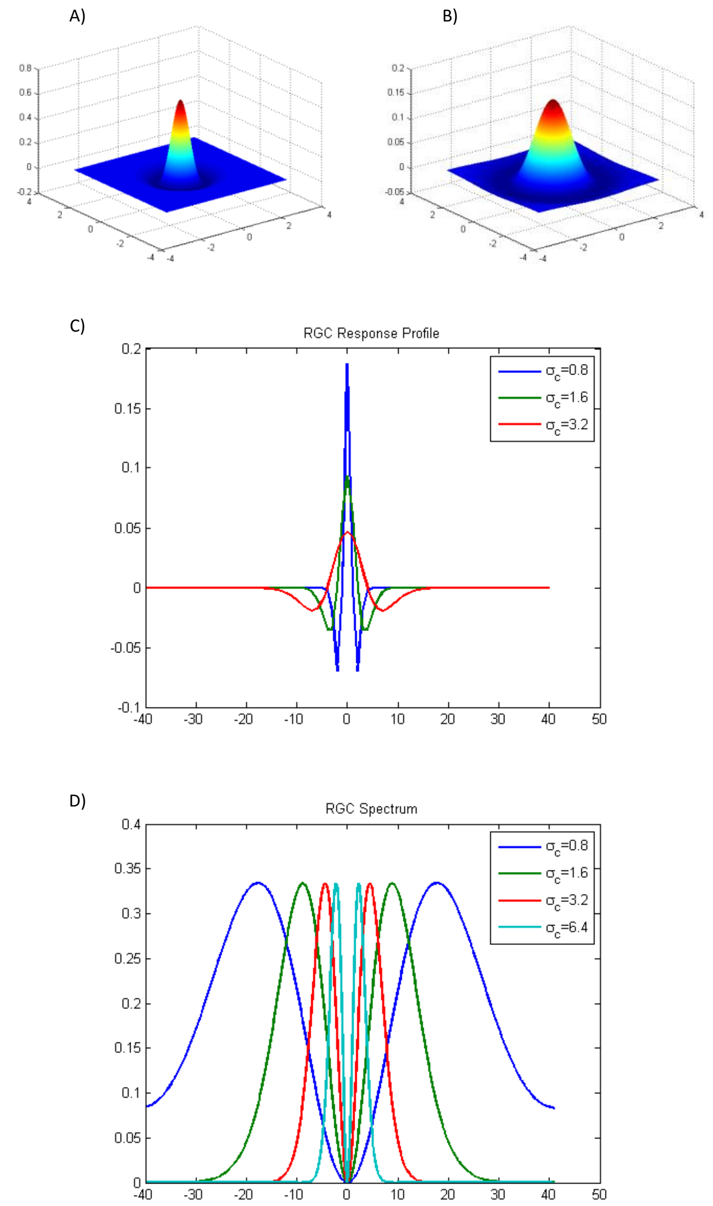


Figure 3-13 Response patterns of Retinal Ganglion Cell spatial receptive fields. (A) and (B) illustrate an example of MGC and PGC responses in spatial domain. (C) and (D) compare cell response profile in spatial domain and frequency domain respectively. As σ increases, the spatial region of RF extends, and its frequency selection becomes lower.

$$f_{LGN}^{spatial}(x,y) = \frac{1}{2\pi\sigma_{LGNc}^2} e^{-\frac{x^2+y^2}{2\sigma_{LGNc}^2}} - \frac{1}{2\pi\sigma_{LGNs}^2} e^{-\frac{x^2+y^2}{2\sigma_{LGNs}^2}}, \tag{3-10}$$

The temporal response pattern (Figure 3-14 C) of modelled M-cells and modelled P-cells are determined as an impulse response filter:

$$f_{LGN}^{temporal}(t) = \frac{c_1 \mathrm{T}(t)}{\sigma_{LGNt1}^2} e^{-\frac{t}{\sigma_{LGNt1}}} - \frac{\mathrm{T}(t)}{\sigma_{LGNt2}^2} e^{-\frac{t}{\sigma_{LGNt2}}}, \tag{3-11}$$

where $\mathrm{T}(t) = 1$ iff $t > 0;$ this forces $f_{LGN}^{temporal}(t) = 0$ when $t \leq 0.$

3.2.1.3 The Primary Visual Cortex (V1)

The primary visual cortex (V1) is the entrance to the visual cortical hierarchy and a main visual processing site where neurons have strong orientation preferences. The modelled V1 has two layers of modelled cells. The first layer contains modelled simple cells (SCs), which receive feed-forward input from modelled LGN cells (modelled M-cells and modelled P-cells). The second layer contains modelled complex cells (CCs), which provide further analysis based on output of model SCs.

[SC] Model V1 Simple Cell

The spatial frequency selectivity of modelled simple cells can be described as a two-dimensional Gabor orientation filter (Daugman 1980, Jones and Palmer 1987) as:

$$f_{V1ss}^{spatial}(x,y) = e^{-\frac{x'^2+\gamma^2 y'^2}{2\sigma_{V1sc}^2}} \cos(2\pi\frac{x'}{\lambda} + \psi), \tag{3-12}$$

where $x' = x\cos\theta + y\sin\theta$ and $y' = y\cos\theta - x\sin\theta$ are the rotational factors, θ denotes the spatial orientation, σ_{V1sc} denotes the sigma of the Gaussian envelope, λ denotes the wavelength of the sinusoidal parameter, and ψ denotes the phase shift. In the current work, $\lambda = \sigma_{V1sc}/0.56$ is used, which yields approximately one-octave bandwidth (Figure 3-15).

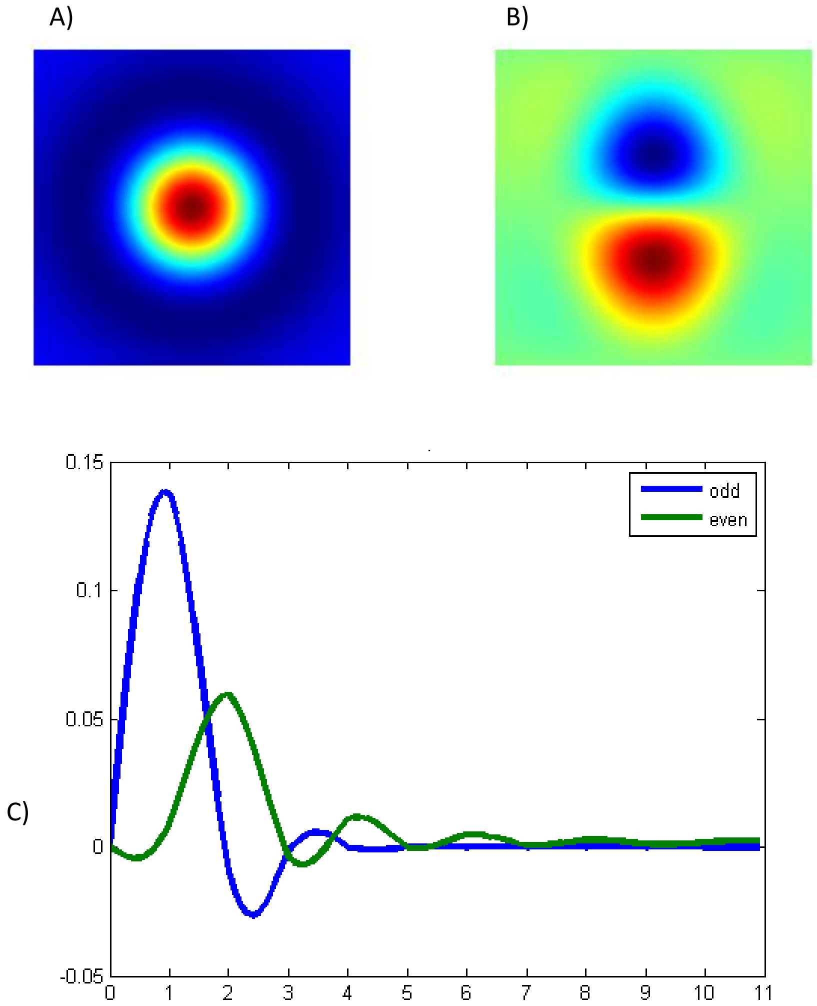


Figure 3-14 LGN spatiotemporal receptive fields. (A) and (B) illustrate spatial response patterns of the LGN cell, 2D views of the real and imaginary parts. Red indicates positive value and blue indicates negative values. Plot C illustrates the temporal response pattern, which is defined as an impulse response filter. The horizontal axis is time, and the vertical axis is the filter value. (C) shows a filter pair (odd and even components).

The temporal frequency profile of modelled SCs is defined by a one-dimensional low-pass filter. The operation is expressed as:

$$f_{V1ss}^{temporal}(t) = \frac{\mathrm{T}(t)}{2\pi\sigma_{V1t}^2} e^{-\frac{t^2}{2\sigma_{V1t}^2}}, \tag{3-13}$$

where $\mathrm{T}(t) = 1$ iff $t > 0$; this forces $f_{V1ss}^{temporal}(t) = 0$ when $t \leq 0$.

SCs that receive signals from M-cells have higher temporal and lower spatial frequency response profiles, where SCs that receive signals from P-cells have lower temporal and band-pass spatial frequency response profiles (Tootell et al. 1988). The spatiotemporal frequency selection is set as follows:

Figure 3-15 illustrates the V1 receptive fields. The proposed usage of Gabor filters leads to even component (plot A) and odd component (plot B) of V1 simple cell. RFs from small to large correspond to σ_{V1sc} from 2.0 to 16.0 respectively as suggested in Table 3-1. The frequency spectrum in 2D (plot C) and 1D (plot D) illustrates the selectivity of these filters in the frequency domain. It is shown that larger σ_{V1sc} values yield lower central frequencies, and vice versa.

[CC] Model V1 Complex Cell

Modelled complex cells integrate simple cell responses. A quadrature pair is used to model the complex cell, which computes the square root over outputs of two simple cells that are 90 degrees out of phase (Figure 3-16). The general complex cell spatiotemporal response can be represented as:

$$R'^{spatiotemporal}_{V1cc} = \sqrt{(R^{even}_{V1ss})^2 + (R^{odd}_{V1ss})^2}\,, \tag{3-14}$$

Given the separate spatial and temporal LGN and simple cell representation, the general spatiotemporal representation along the feed-forward pathway from LGN to V1 complex cell

(Figure 3-17) can then be derived following (Adelson and Bergen 1985):

$$R_{V1cc}^{spatiotemporal} = \Theta\left[R_{V1ss}^{spatial_{odd}} R_{V1ss}^{temporal_{even}} - R_{V1ss}^{spatial_{even}} R_{V1ss}^{temporal_{even}}\right], \quad (3\text{-}15)$$

3.2.1.4 The Visual Cortex V2

Modelled V2 neurons receive feed-forward input from Modelled V1 CCs. Cells in area V2 have larger receptive fields than that of V1, and with end-stopped response patterns. Inspired by (Dobbins et al. 1987, Boynton and Hegde 2004), four types of V2 end-stopped receptive fields have been proposed in (Rodriguez-Sanchez and Tsotsos 2011). End-stopped cells are modelled as a V1 simple cell suppressed by two displaced V1 complex cells at elongated ends. Suppression allows end-stopped cells to perform curvature estimation. Depending on the rotational components, the end-stopped cell computes convexity/concavity. In this work, Type 4 cells (one of the four end-stopped cells, the most complicated one in terms of computational representation) are implemented:

$$R_{V2}^{end-stopped} = \Theta\left[c_0 R_{V1ss}^{center_{\theta 0}} - \left(c_1 R_{V1cc}^{displaced_{\theta 1}} + c_2 R_{V1cc}^{displaced_{\theta 2}}\right)\right], \quad (3\text{-}16)$$

where $R_{V1ss}^{center_{\theta 0}}$ is the center SC response, $R_{V1cc}^{displaced_{\theta 1}}$ and $R_{V1cc}^{displaced_{\theta 2}}$ are two displaced CC responses, θ_1 and θ_2 are the relative orientations from the center SC orientation θ_0. Parameters c_0, c_1, and c_2 are weighting constants. Figure 3-18 shows an example of a Type 4 end-stopped cell with two displaced CCs of 45° relative to the center SC.

Table 3-3 σ_{V1sc} set to simulate V1 Simple Cells.

σ_{V1sc} for dorsal V1 neuron	2.0, 4.0, 8.0
σ_{V1sc} for ventral V1 neuron	4.0, 8.0, 16.0

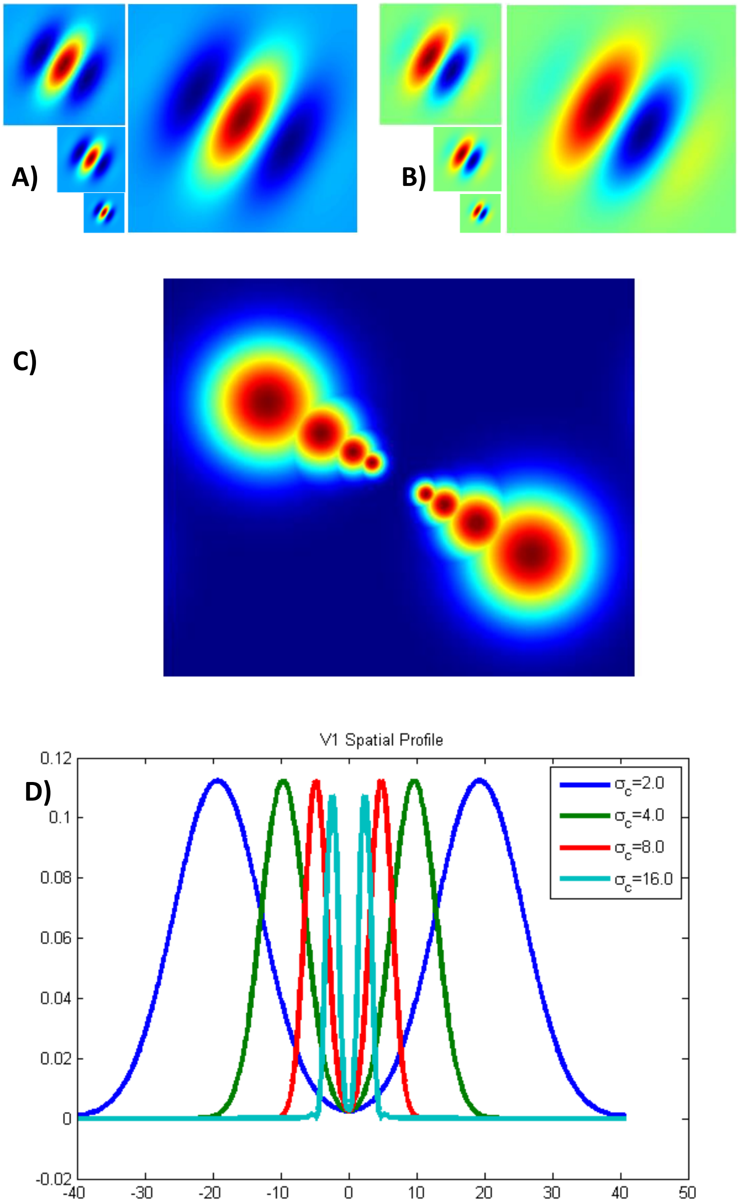


Figure 3-15 A V1 simple cell spatial receptive field. (A) and (B) illustrate even and odd component of the proposed usage of Gabor filters. RFs from small to large correspond to σ values from 2.0 to 16.0 respectively as suggested in Table 3-1. The frequency spectrum in 2D (C) and 1D (D) illustrates the selectivity of these filters in the frequency domain. It is shown that larger σ value yields lower central frequency, and vice versa.

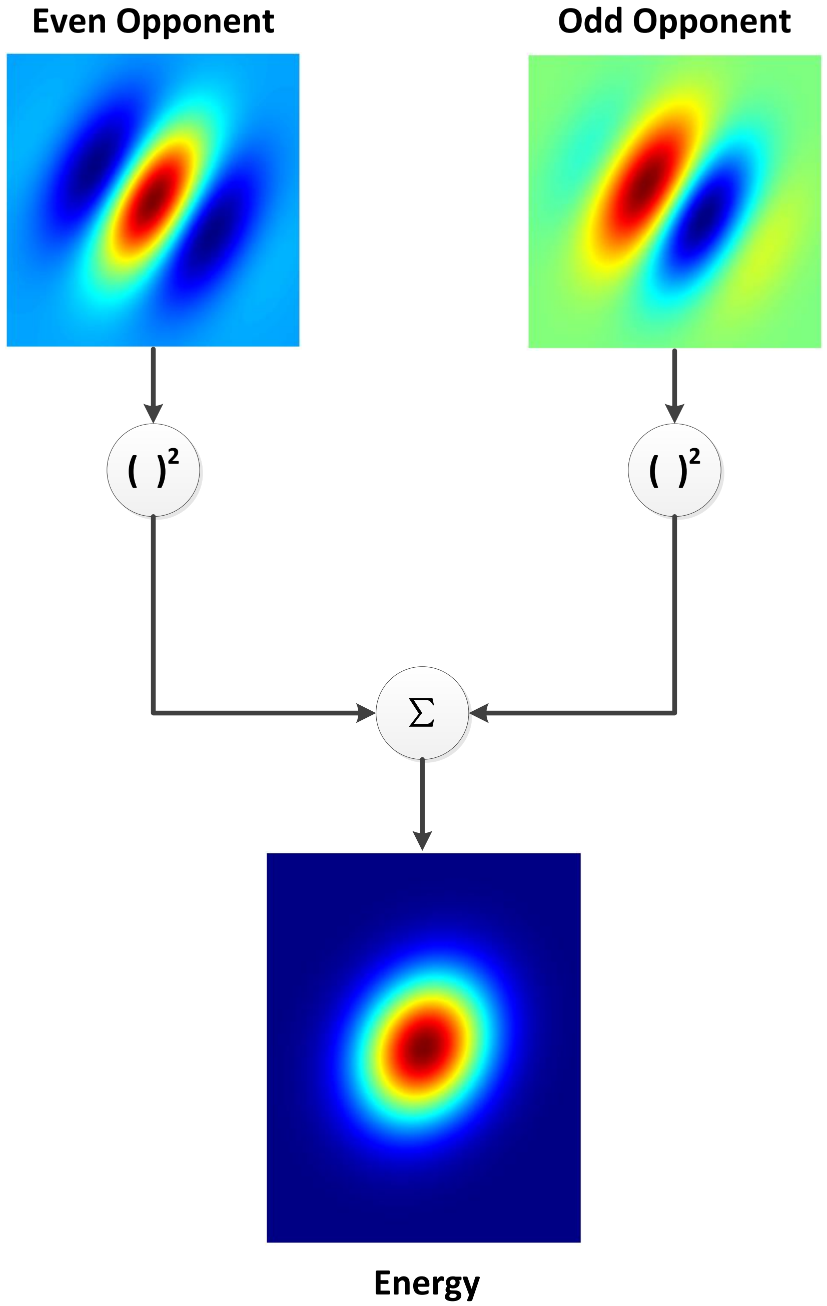


Figure 3-16 The formation of a V1 complex cell receptive field. A quadrature pair is used to model the complex cell. This is from an energy summation perspective, which computes the square root over outputs of two simple cells that are 90 degrees out of phase.

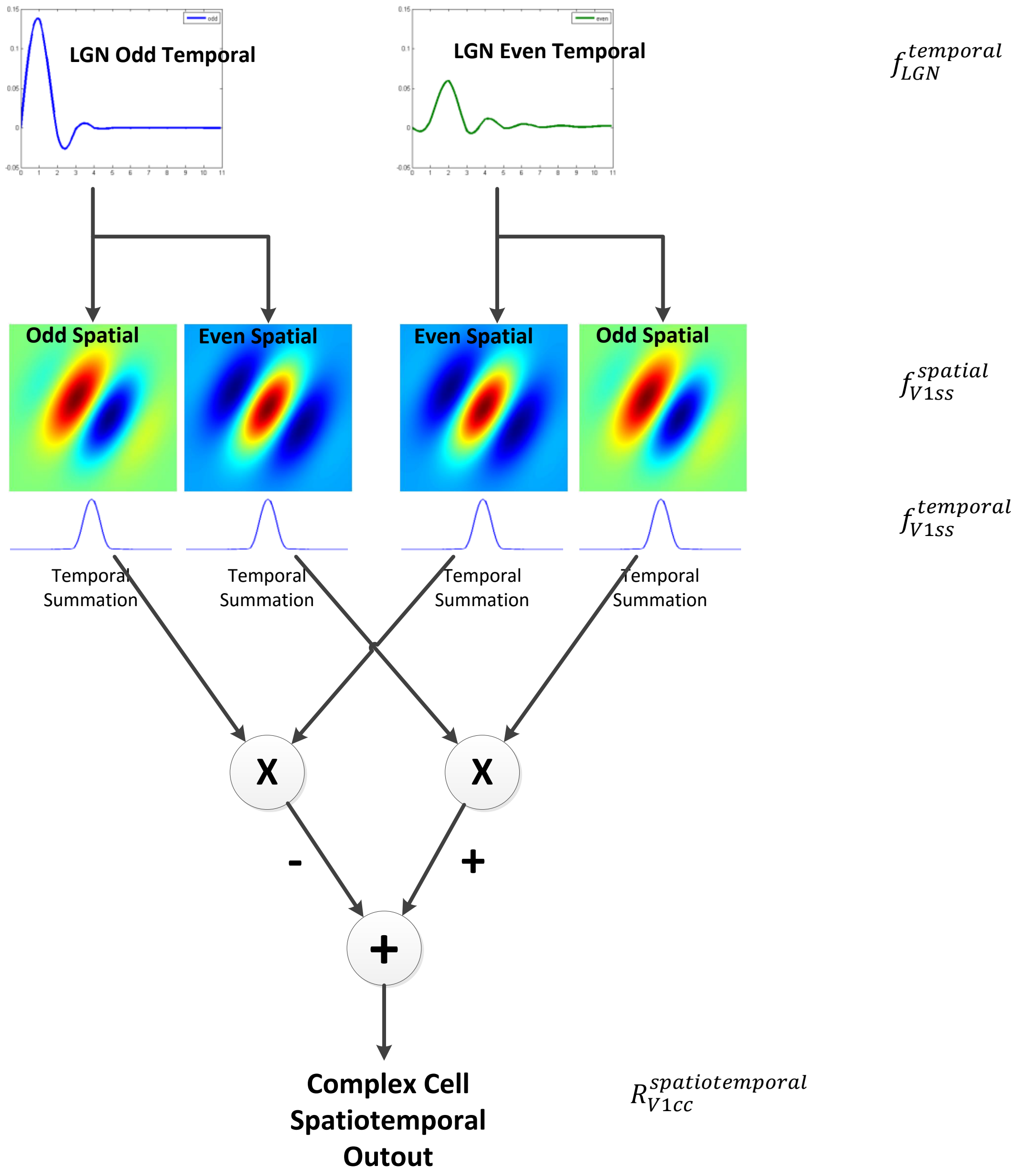


Figure 3-17 LGN-V1 spatiotemporal analyses. Given the separate spatial and temporal LGN and simple cell representation, the general spatiotemporal representation along the feed-forward pathway from LGN to V1 complex cell can then be derived following (Adelson and Bergen 1985). Circles with "X" indicates dot multiple operation, and circles with "+" indicates plus operation.

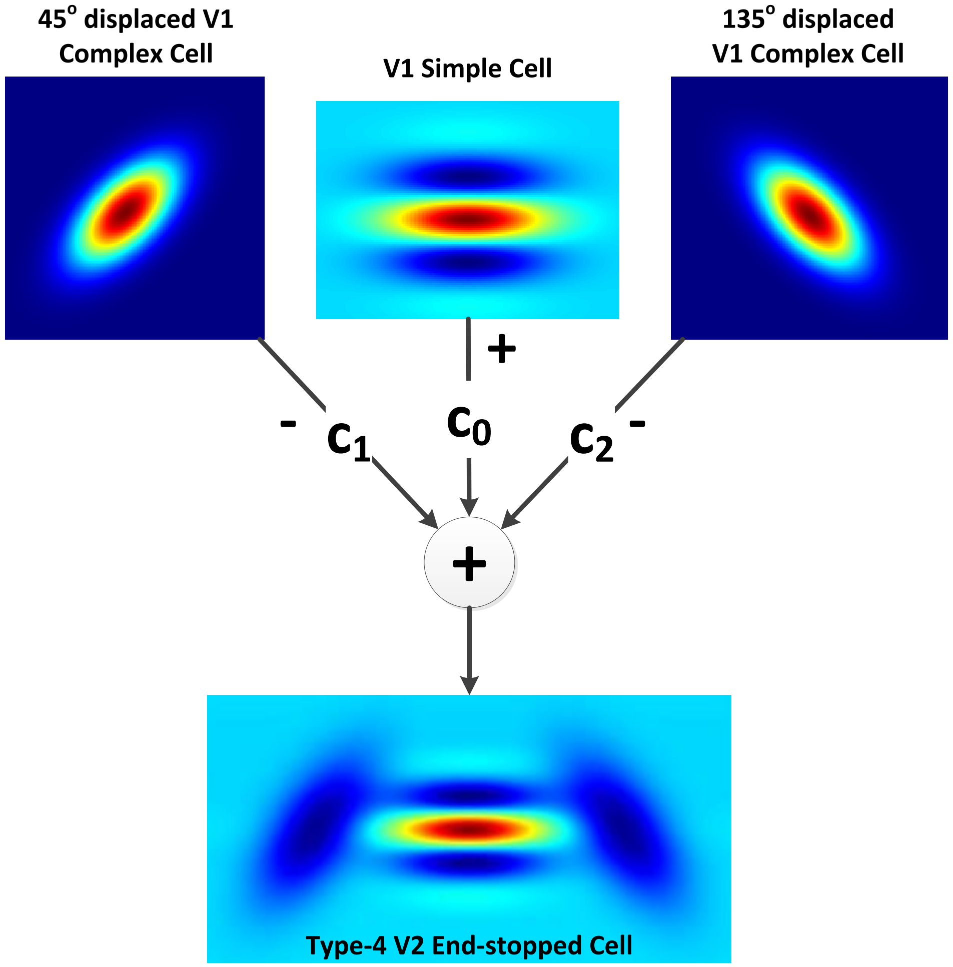

45° displaced V1 Complex Cell
V1 Simple Cell
135° displaced V1 Complex Cell
+
-
c_1
c_0
c_2
-
+
Type-4 V2 End-stopped Cell

Figure 3-18 The formation of a Type-4 V2 end-stopped cell. It calculates the difference between a simple cell and two displaced complex cells with rotated component. Compared with Type-3 cells it can distinguish between curvature directions. C0, C1 and C2 represent weightings to add the components.

3.2.1.5 The Middle Temporal Cortex (MT)

Area MT is an important cortical region in the dorsal pathway, which processes local motion. It has larger receptive fields than area V1 and is sensitive to high-temporal-low-spatial frequency moving patterns. This spatiotemporal detection characteristic of MT neurons has been shown to respond to both velocity changes (Weller et al. 1984) and velocity gradients (Treue and Andersen 1996) . We follow (Tsotsos et al. 2005) to separately model these two types of MT neurons.

Modelled MT neurons receive and integrate feed-forward input from the V1 CCs. Modelled velocity-change neurons (MTvc) accumulate opponent energy from V1 neurons within a larger spatial area. The spatiotemporal property of MTvc is determined by a Gaussian envelope as:

$$f_{MTvc}^{spatiotemporal}(x,y,t) = e^{-\frac{x^2+\gamma_{MTvc}^2 y^2+\xi_{MTvc}^2 t^2}{2\sigma_{MTvc}^2}}, \qquad (3\text{-}17)$$

where σ_{MTvc} denotes the Gaussian standard deviation, which defines the spatial extend of the receptive field. γ_{MTvc} is the spatial aspect ratio, where $\gamma = 1$ yields a circular MT spatial receptive field. ξ_{MTvc} is the temporal aspect ratio.

Modelled velocity-gradient neurons (MTvg) detect changes in velocity across V1 neurons within a larger spatial area. Depending on the velocity direction and velocity gradient, (Treue and Andersen 1996) defines four gradient types: clockwise shear, counter-clockwise shear, stretching and compression. Figure 3-19 shows a polarized view to illustrate these four types. If direction of motion and direction of gradient has a right angle, then it is a shearing motion. If direction of motion and direction of gradient is consistent, then it is a stretching motion. If direction of motion and direction of gradient oppose each other, then it is a compression motion.

To facilitate the discussion, we separate the spatial and temporal RF properties. The temporal property of MTvg is determined by a Gaussian envelope as:

$$f_{MTvg}^{temporal}(t) = e^{-\frac{\xi_{MTvg}^{2} t^{2}}{2\sigma_{MTvg}^{2}}}, \tag{3-18}$$

where σ_{MTvg} defines the spread on the temporal axis, and ξ_{MTvg} is the temporal aspect ratio.

Spatially, a modelled MTvg accumulates V1 speed gradients. To formalize this property, $(x - y - s)$ columnized representation of V1 opponent responses is assumed, where x and y represent the spatial location, and s denotes spatiotemporal orientations or velocity, i.e. s =1, 2, 3 denotes slow / middle / fast speeds respectively. The summation of MTvg is modelled as template matching on spatiotemporal planes:

$$R_{MTvg}^{spatial}(x, y) = \sum_{s \in 1,2,3} R_{V1cc}^{s}(x, y) \circledast T_s(x, y), \tag{3-19}$$

Where $\circledast$ denotes convolution, $\mathrm{R}_{\mathrm{V1cc}}^{\mathrm{s}}(\mathrm{x}, \mathrm{y})$ denotes V1 complex-cell response to velocity s. T_{s} is the template for velocity s.

In our implementation, a template matching set contains three 5-by-5 templates, with each template representing a speed. Table 3-4 shows an example of a set of templates to detect gradient changes of rightward stretching. Intuitively, when a stimulus is stretching to the right, one would expect to observe strong energy shift, spatially from left to right, and speed-wise from slower to faster. This can be revealed in Table 3-4 by 1s on the left-most column in low-speed motion template, 1s on the middle column in middle-speed motion template, and 1s on the right-most columns in high-speed motion template.

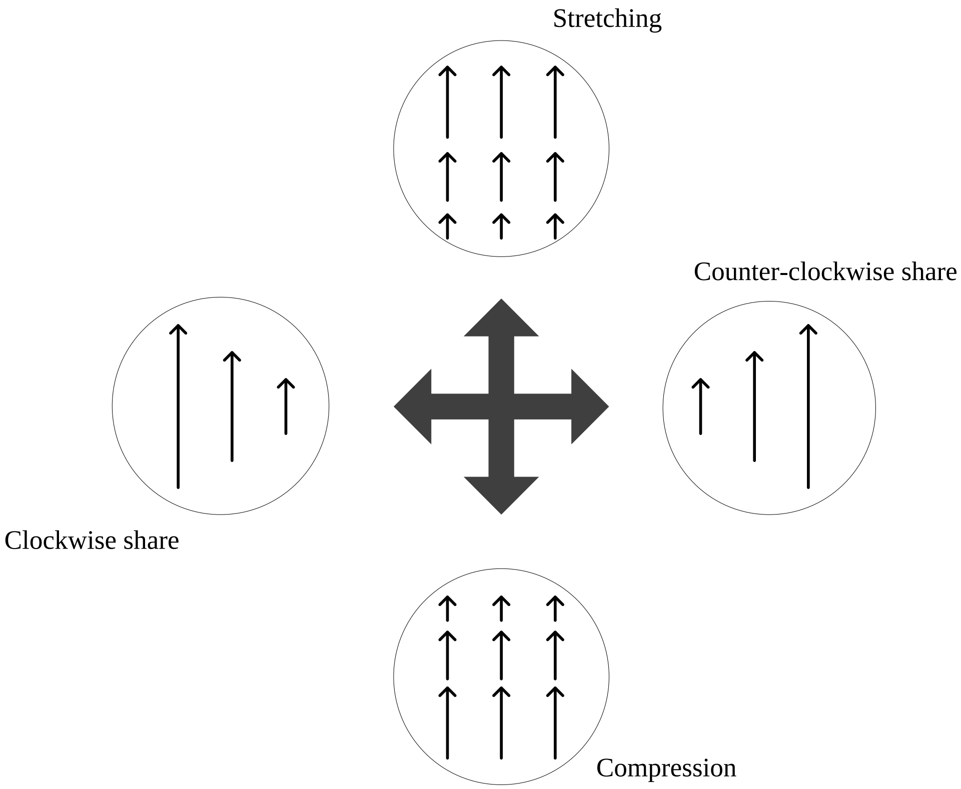


Figure 3-19 Basic gradient types defined gradients (Treue and Andersen 1996). Depending on the velocity direction and velocity gradient, four gradient types are defined: clockwise shear, counter-clockwise shear, stretching and compression.

Table 3-4 An example a set of templates to detect gradient changes of rightward acceleration. When a stimulus moves toward right with increasing speed, one would observe strong energy shift, spatially from left to right, and speed-wise from slower to faster. This can be revealed by 1s on the left-most column in low-speed motion template, 1s on the middle column in middle-speed motion template, and 1s on the right-most columns in high-speed motion template.

1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0
1.0	0.0	0.0	0.0	0.0

s = 1 (Low-speed motion)

0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0
0.0	0.0	1.0	0.0	0.0

s = 2 (Mid-speed motion)

0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	1.0
0.0	0.0	0.0	0.0	1.0

s = 3 (High-speed motion)

3.2.1.6 The Medial Superior Temporal Cortex (MST)

Neurons of area MST have larger receptive fields than the MT neurons, and are tuned to more complex motion patterns: expand or approach, contract or recede, and rotation (Duffy and Wurtz 1991, Duffy and Wurtz 1997, Duffy 1998, Smith et al. 2006). In our model, a modelled MST neuron further integrates activations from the modelled MT neurons. The spatiotemporal properties of modelled MST neuron can be defined as:

$$f_{MST}^{spatiotemporal}(x,y,t) = e^{-\frac{x^2+\gamma_{MST}^2 y^2+\xi_{MST}^2 t^2}{2\sigma_{MST}^2}}, \qquad (3\text{-}20)$$

where σ_{MST} denotes the Gaussian standard deviation, which defines the spatial extend of the receptive field, γ_{MST} is the spatial aspect ratio, where $\gamma = 1$ yields a circular MT spatial receptive field, and ξ_{MST} is the temporal aspect ratio.

3.2.1.7 A Summary of the Modelled Visual Hierarchy

The formalized visual hierarchy includes a representation of both the dorsal pathway and the ventral pathway. From the feature processing point of view, the two pathways interpret different spatiotemporal visual characteristics. We modelled the feed-forward processing along the visual hierarchy using a set of image filters, which are summarized in Table 3-5.

Table 3-5 A summary of the modelled hierarchy.

Visual area	Cell type	Spatial filter	Spatial σ	Temporal filter
RGC	midget cell	DoG	0.8, 1.6	Low-pass
	parasol cell	DoG	1.6, 3.2	Low-pass
LGN	parvocellular cell	DoG	1.6	Low-pass
	magnocellular cell	DoG	2.0	Impulse Response Filter
V1	Simple cell	Log-Gabor	ventral: 2.0, 4.0, 8.0 dorsal: 4.0, 8.0, 16.0	Low-pass
	Complex cell	Energy summation	ventral: 4.0, 8.0, 16.0 dorsal: 8.0, 16.0, 32.0	Low-pass
V2	Type-4 cell	Energy summation	8.0, 16.0, 32.0	Low-pass
MT	velocity cell	Energy summation	32.0	Low-pass
	gradient cell	Energy summation	32.0	Low-pass
MST	MST cell	Energy summation	64.0	Low-pass

3.2.2 Formalization of Early Recurrence

3.2.2.1 General Formula

Early recurrence from the dorsal pathway modulates ventral processing. The modulation inhibits ventral neural response representation. The general modulation processing can be expressed as:

$$R'_{ventral} = \Theta[R_{ventral} \cdot Inh_{dorsal}], \tag{3-21}$$

where $R_{ventral}$ denotes the original ventral representation, which is driven by feed-forward signal. · denotes the recurrent operation. Inh_{dorsal} denotes the recurrent representation that is generated from the dorsal pathway. Θ is rectification function. The early recurrent representation Inh_{dorsal} is further defined as:

$$Inh_{dorsal} = \frac{\omega_i R_{dorsal\_i}}{\left\|\sum_j \omega_j R_{dorsal\_j}\right\|}, \tag{3-22}$$

where R_{dorsal_i} denotes the dorsal neuron response to preferred spatiotemporal orientation i. This is the result of feed-forward computation. ω_i denotes its weighting strength, with $\omega_i \in [0, 1]$. Further, the denominator $\left\|\sum_j \omega_j R_{dorsal\_j}\right\|$ indicates that the recurrent representation is normalized. The modulation reaches its stable state when $R'_{ventral} = R_{ventral}$. Since the feedback representation is non-zero, the model implies that modulation will be stabilized when the system lacks motion cues, or the motion cue is stopped (motion is off-set).

Depending on the setting of ω_i, one may derive the three different modulation patterns proposed in the previous section:

Pattern-1 (Anisotropic): early recurrence has a spatially anisotropic suppression pattern. Each ventral neuron receives recurrent signals from dorsal neurons of the same spatiotemporal preference. This is done by:

$$\omega_i = \begin{cases} 1, if\ orientation\ is\ also\ preferred\ by\ venral\ neuron \\ 0, otherwise \end{cases}, \tag{3-23}$$

Pattern-2 (Isotropic): the ventral neuron receives recurrent signals from all dorsal neurons in an equal manner. To do this, one can fix $\omega_i = 1$ for all dorsal neurons. This isotropic representation is an equal summation of dorsal strengths.

Pattern-3 (General): In a more generalized manner, the ventral neuron receives recurrent signals of different strengths. Each ventral neuron has a preferred tuning profile. Maximum suppression is achieved if the recurrent signal matches the preferred profile and least affected at its non-preferred profile. Between the preferred and non-preferred profile, the suppressiveness may show linear or non-linear tuning effects.

3.2.2.2 Recurrence between Area MT and Area V1/V2

Recurrence between area MT and area V1/V2 takes place between dorsal area MT and ventral layers of area V1 and area V2. The recurrence applies the MT representation to modulate V1 edge representation, and V2 curvature representation. Specifically, given the aforementioned V1 response ($R_{V1cc}^{spatiotemporal}$), V2 response ($R_{V2}^{end-stopped}$) and MT response, the general early recurrent equation is then:

$$R_{V1cc-MT}^{\prime\, spatiotemporal} = \Theta\left[R_{V1cc}^{spatiotemporal} \cdot Inh_{MT}\right], \tag{3-24}$$

$$R_{V2-MT}^{\prime\, end-stopped} = \Theta\left[R_{V2}^{end-stopped} \cdot Inh_{MT}\right], \tag{3-25}$$

where Inh_{MT} is a combined representation of $\mathrm{R_{MTvc}}$ and $\mathrm{R_{MTvg}}$. $\mathrm{R_{MTvc}}$ indicates directional features and $\mathrm{R_{MTvg}}$ represents velocity gradients:

$$Inh_{MT} = \frac{\omega_i(R_{MTvc\_i} + R_{MTvg\_i})}{\left\|\sum_j \omega_j(R_{MTvc\_j} + R_{MTvg\_j})\right\|}, \tag{3-26}$$

3.2.2.3 Recurrence between Area MST and Area V2

Recurrence between area MST and area V2 modulates V2 curvature representation. Specifically, given the aforementioned V2 ($R_{V2}^{end-stopped}$) response and MST responses ($\mathrm{R_{MST}}$), the general early recurrent equation is then:

$$R_{V2-MST}^{\prime\, end-stopped} = \Theta\left[R_{V2}^{end-stopped} \cdot Inh_{MST}\right], \tag{3-27}$$

where Inh_{MST} denotes the MST recurrent representation:

$$Inh_{MST} = \frac{\omega_i R_{MST\_i}}{\left\|\sum_j \omega_j (R_{MST\_j})\right\|}, \tag{3-28}$$

3.2.3 A Full Feed-forward Hierarchy Example

The model has been implemented and investigated in several scenarios to address different computer vision algorithms. Both synthetic and real images have been used to study the impact of early recurrence to lower-level visual feature processing.

Before discussing the impact of early recurrence, it is important to move one step further on these feed-forward pathways to illustrate how stimuli are computed and interpreted through the hierarchy. In what follows, we use synthetic octagon stimuli with complex motion patterns as an example.

Figure 3-20 shows an example using purely motion-defined objects. Two octagons with plaid patterns are used. The left octagon rotates in a clockwise direction and the octagon on the right rotates in a counter-clockwise direction. The two objects move at equal speed.

Figure 3-21 shows the response of retinal ganglion cells. The top row shows the on-center and off-center PGC responses. Since PGCs have larger receptive fields, they respond to coarse level intensity variations. The bottom row shows on-center and off-center MGC responses. Due to the small receptive fields, PGCs respond to fine-level intensity changes.

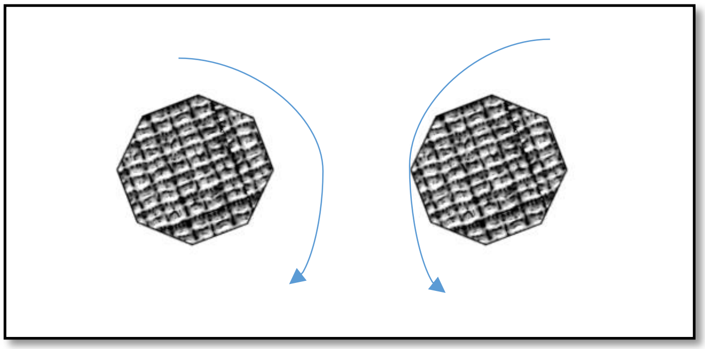


Figure 3-20 Input stimuli are two octagons with plaid patterns are used. The octagon on the left rotates clockwise and the octagon on the right rotates counter-clockwise.

Figure 3-22 shows output of the LGN spatiotemporal responses. The top row shows the parvocellular cell responses. Since parvocellular cells are not sensitive to temporal variations, the output of parvocellular cells accumulates Midget Retina Cells responses with a low-pass spatiotemporal filtering effect.

Row 2 to row 4 of Figure 3-22 illustrates the magnocellular cell responses of the LGN. Since the magnocellular cells are very sensitive to temporal variations, three types of magnocellular cells are implemented. They respond to fast, middle and slow temporal variation respectively. We follow (Adelson and Bergen 1985) for the spatiotemporal analysis to derive odd and even cell responses. For each magnocellular cell, the left image shows the even cell responses and the right image shows the odd cell responses. We compare among the three cell types. It illustrates that as temporal resolution goes from fast (row 2) to slow (row 4), responses to the surround region of the octagons are gradually weakened, while the responses to the center region of the octagons become stronger. This is consistent with the stimuli that since the two octagons are rotating, the speed near the centre of the octagon is slower than in the surrounding region.

Figures 3.23 to 3.25 illustrate area V1 responses to the magnocellular cell inputs. V1 cells extract different spatiotemporal orientations. To simplify the computation, the current study separates spatiotemporal analysis into spatial and temporal analysis. These figures illustrate how V1 extracts fast-, middle-, and slow-speed motion input into 12 orientations respectively.

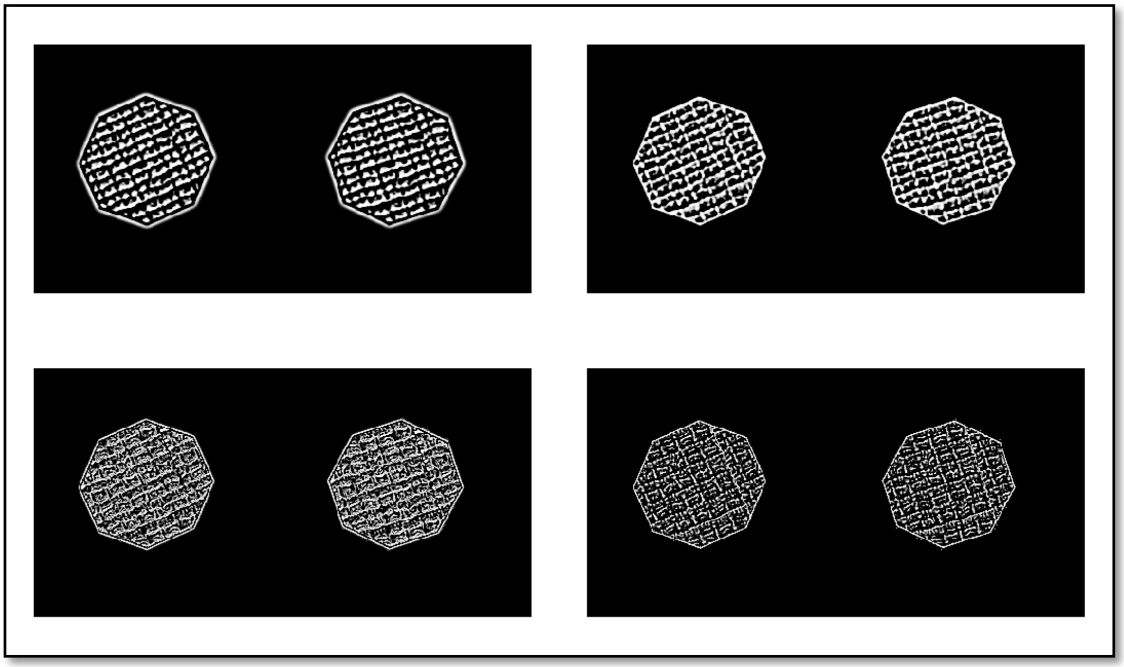


Figure 3-21 The RGC response. Top row: parasol on-centre and off-centre responses. Bottom row: midget cell on-centre and off-centre responses.

Figure 3-26 compares fast-, middle- and slow-speed motion patterns with motion directed to the left. The speed at the outer region of the octagons is faster than that at the center region. This observation is consistent with the patterns shown here: there are significant responses at the outer regions in the fast-speed cell response (top-row), and stronger responses at the center regions in the slow-speed cell response (bottom-row). Responses of the middle-speed cell (middle-row) highlight the regions in between.

Figure 3-27 illustrates area V1 responses to parvocellular cell input.

Figure 3-28 illustrates area MT responses. Area MT accumulates area V1 responses and further extracts complex motion patterns. The figure clearly illustrates that the two octagons contain strong shearing motion.

Figure 3-29 illustrates area V2 responses to area V1 input. Compared with area V1, three points are observed in the V2 presentation. The first observation is that V2 responses to curvatures (e.g. curved-boundary of the octagons). The second observation is that due to larger receptive fields, area V2 extracts more complete edges. Last but not least, due to the inhibitive nature of the V2 receptive fields, V2 representation clearly highlights edges and curvatures from the noisy patterns.

An interesting aspect of the proposed model lies in the ability of the dorsal pathway to respond to different types of motion patterns and coarse-scale spatial information. These fast-computed representations facilitate ventral processing in the form of local context that the ventral pathway is unable to compute by itself. In the next chapter, we will show how these dorsal presentations modulate edge and curvature representations computed in the ventral pathway.

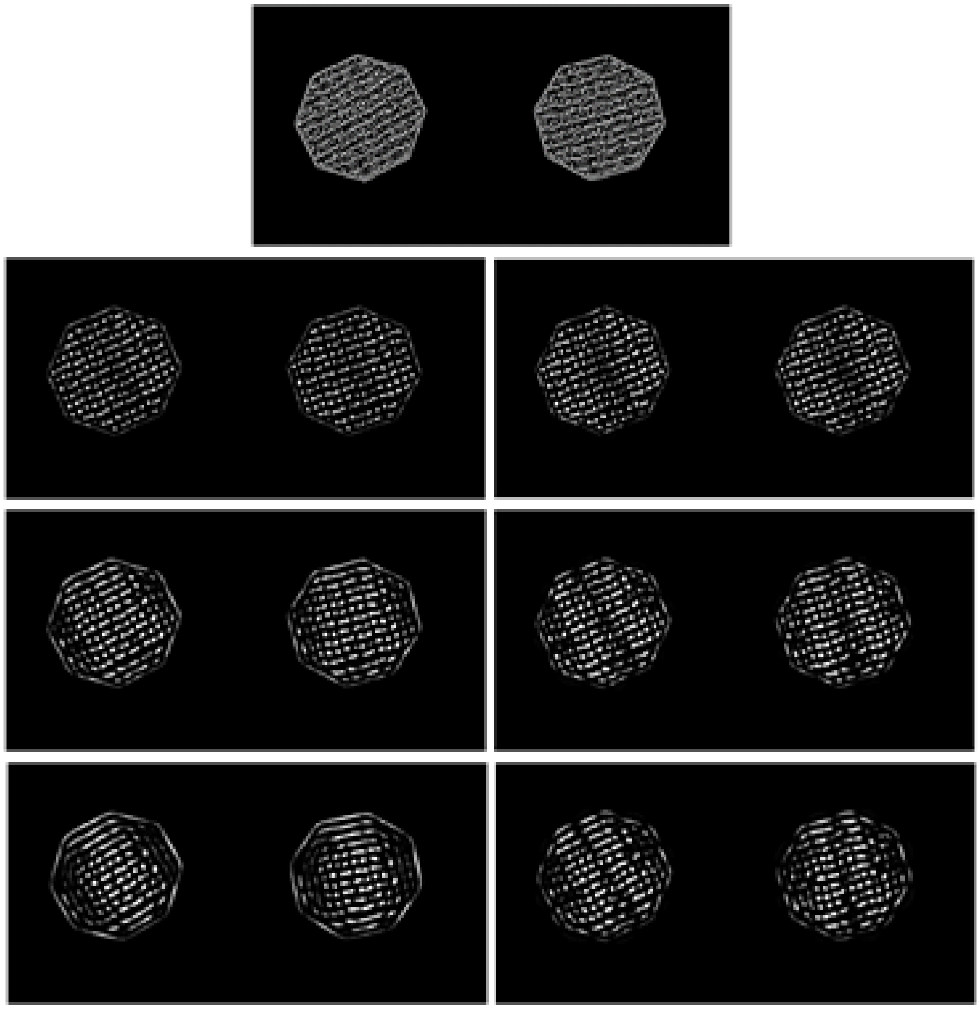


Figure 3-22 The LGN responses to RGC inputs. Top row: parvocellular cell responses. Bottom rows: magnocellular cell responses. Left and right figures in each row illustrate even and odd cells respectively following (Adelson and Bergen 1985). Row 2 shows magnocellular cell response to fast motion; row 3 shows middle-speed motion; and row 4 shows slow speed motion.

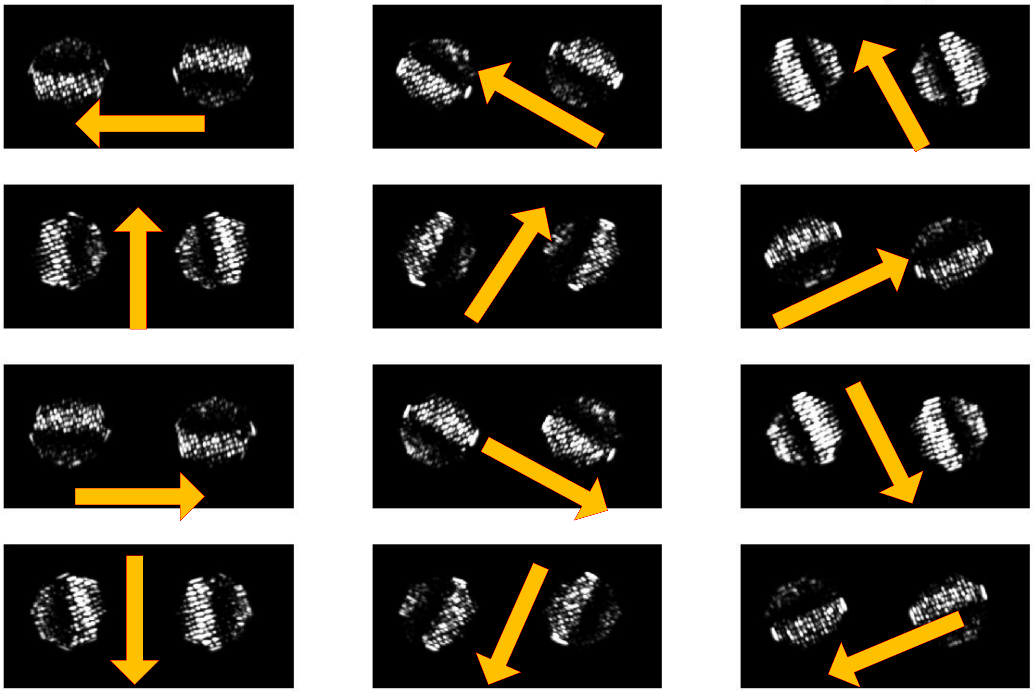


Figure 3-23 V1 responses to fast-speed motion stimuli. Arrows indicate preferred orientations.

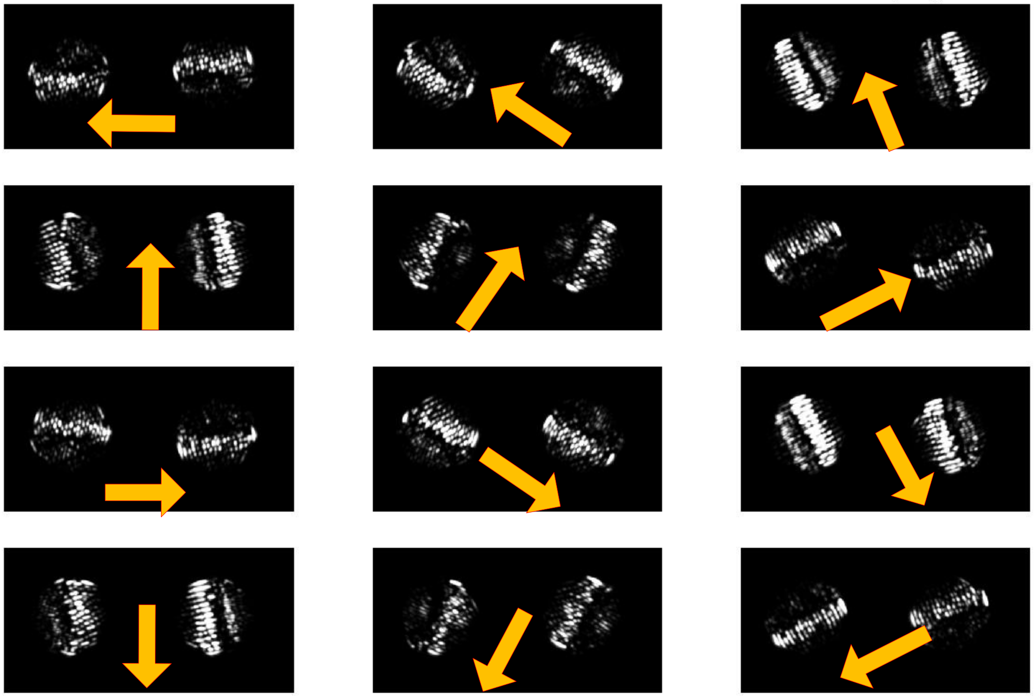


Figure 3-24 V1 responses to middle-speed motion stimuli. Arrows indicate preferred orientations.

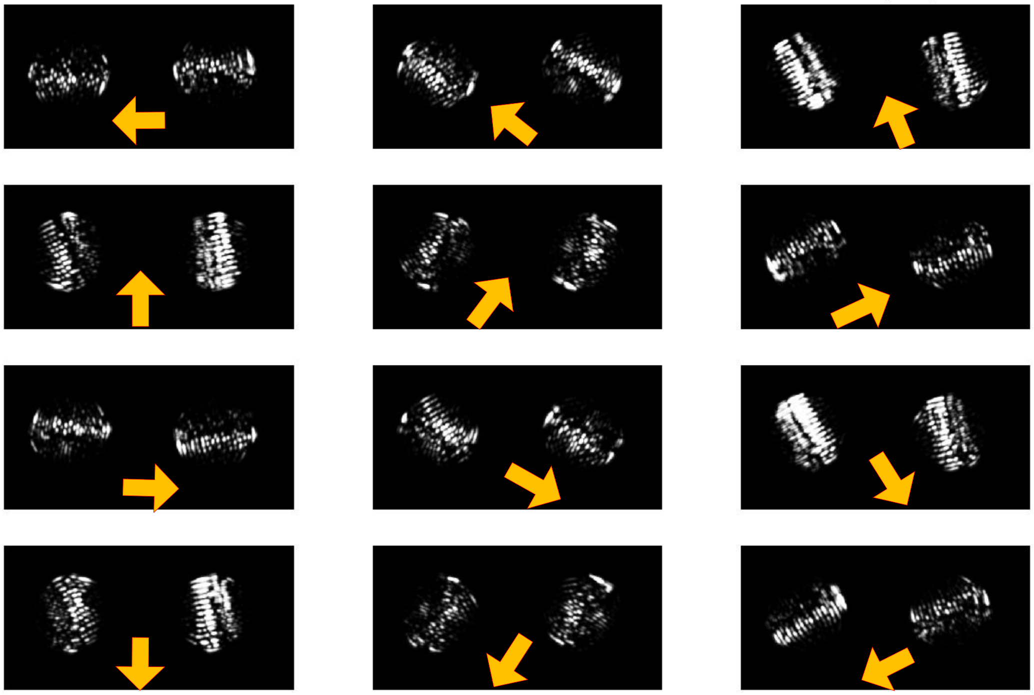


Figure 3-25 V1 responses to slow-speed motion stimuli. Arrows indicate preferred orientations.

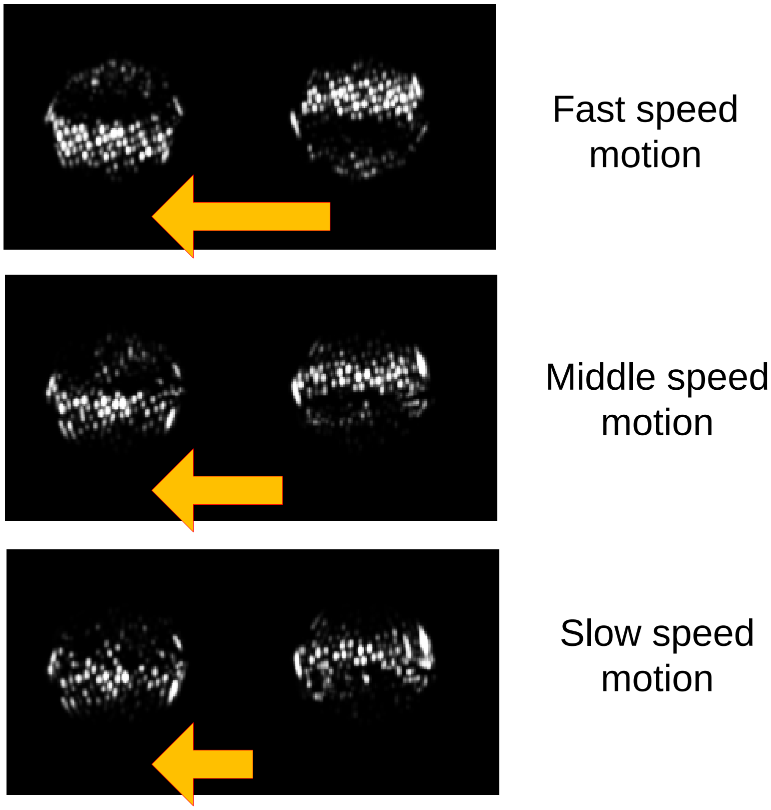


Figure 3-26 V1 responses to fast-, middle- and slow-speed motion patterns with moving direction to the left. Arrows indicate preferred orientations. Length of arrows illustrates speeds.

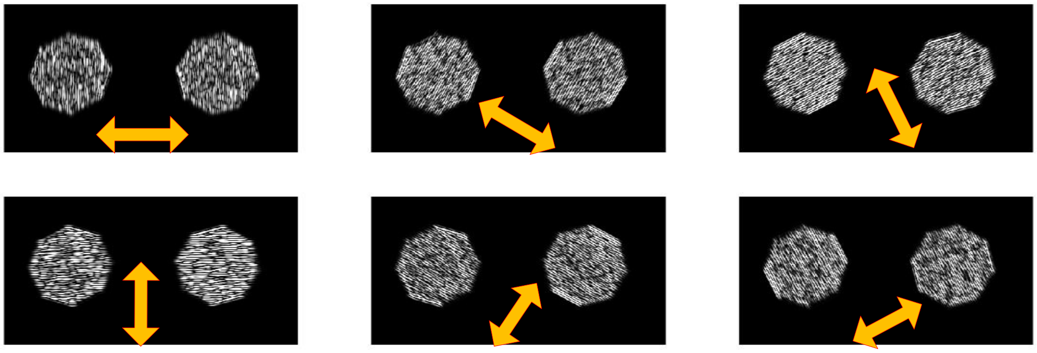


Figure 3-27 V1 responses to parvocellular cell inputs. Arrows indicate preferred orientations.

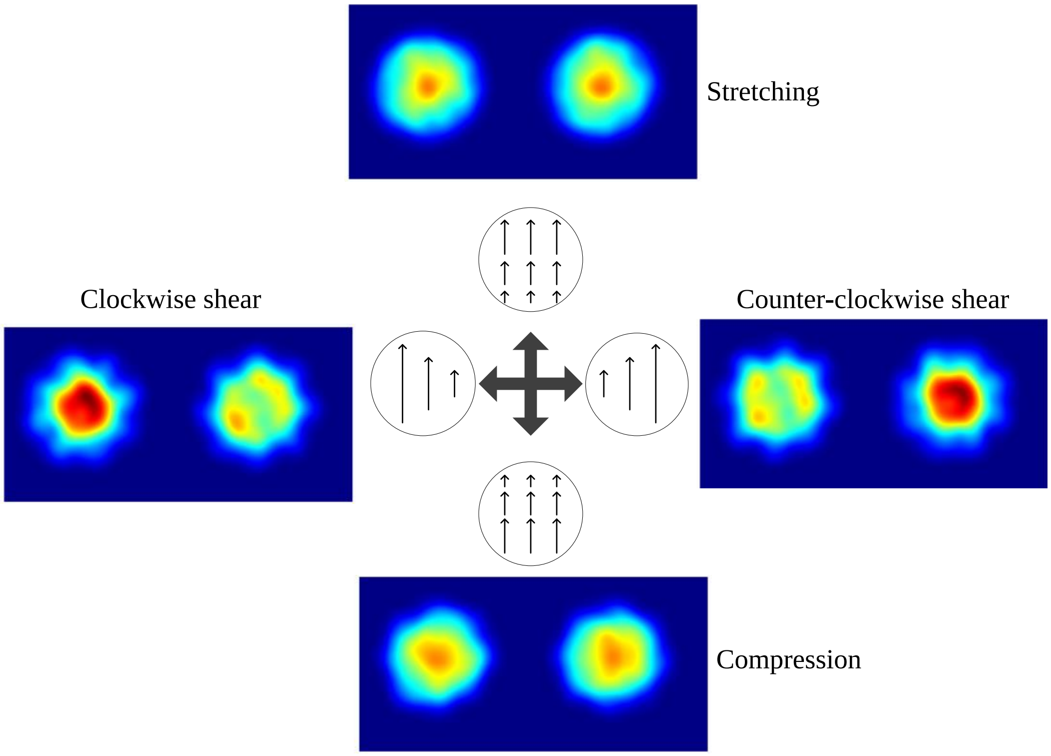


Figure 3-28 MT responses reveal that the two circles contain strong shear motion.

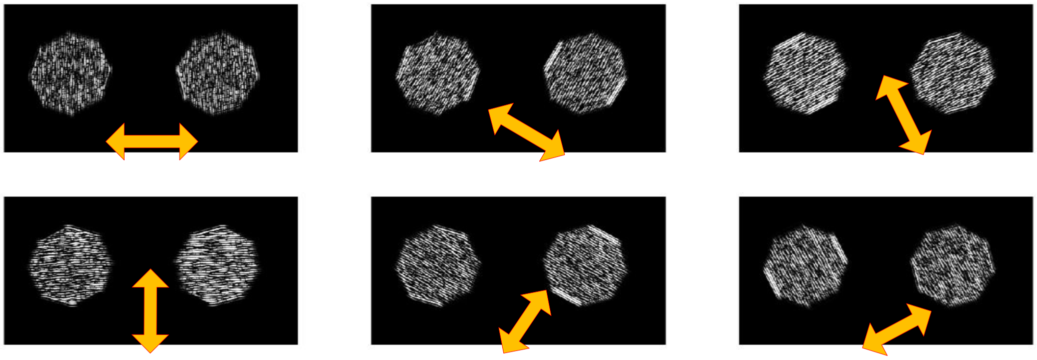


Figure 3-29 V2 responses to curvatures. Arrows indicate preferred orientations.

Chapter 4. Experiments

In Chapter 3, we proposed a computational model of early recurrence. Using synthetic moving octagons, we illustrated how visual stimuli are processed by the proposed model. The ultimate goal of this work is to show how the proposed computational components might improve computer vision algorithms. However, before discussing topics in computer vision, in this chapter we will illustrate the connections between the proposed model and the biological vision.

Specifically, we report our efforts to simulate two experiments on biological subjects: a figure-background segregation experiment on macaques (Hupé et al. 1998) and a Kanisza illusory rectangle experiment on humans (Seghier et al. 2000b). These two experiments are well-known studies of the impact of lower-level recurrence of visual processing. In both experiments, the authors conduct quantitative analysis to measure the extent to which recurrent activity impacts the lower-level ventral response. They drew similar conclusions, which include the verification of the existence of early recurrence and the observation of modulation effect on the ventral processing.

We followed the test protocols from the original manuscripts to set up testing stimuli and experiment conditions. We wanted to know whether the proposed computation could reproduce observations that have been reported on biological subjects. Additionally, to study the recurrent impact, we investigated alternative recurrent operations other than the multiplicative inhibition. If the simulations are successful, the proposed computational model itself will serve as an evidence to support the fast-brain hypothesis (Bullier 2001), and we will also be confident to apply the proposed model in computer vision systems.

4.1 Figure-Background Segmentation

By inactivation of the higher-order visual areas V5/MT of macaques, Bullier and his colleagues studied the hypothesis that feedback amplifies neural activity on lower-level visual areas, such as areas V1, V2, and V3 (Hupé et al. 1998, Bullier et al. 2001). In this experiment, the centre of the visual field contains a bar. Background contains randomly distributed checker-box. The bar has the same width as the background checker-box. The background checker-box has similar appearance to the central bar. By setting the bar and the background in three different relative motion patterns, the authors showed that early feedback representation was important in differentiating the figure from the background. Feedback facilitates response to an object moving within the classical receptive field and enhances suppression evoked by background stimuli in the surrounding region. This facilitatory effect strengthens for stimuli of lower saliency. That is, the modulation effect is prominent when stimuli are of low visibility. The study suggests a lower-level recurrent mechanism that reduces interference of moving distractors, especially when the stimuli are in complicated moving patterns.

4.1.1 Stimuli

The current work simulated this experiment. We generated the same stimuli as those described in (Hupé et al. 1998). The stimuli include a solid bar moving constantly in front of randomly distributed checker-box. In the original experiment, the moving bar was optimized in the approximate size and velocity for the V1 neuron, and the orientation was optimized to within 15° by measurement of an orientation tuning curve. In our experiment, this is done by setting the filter spatiotemporal properties in the approximate size and velocity of the stimuli.

We investigated early recurrence using three sets of relative moving patterns following the original work. In Set 1, the bar moves in front of the stationary checker-box (Figure 4-1a). In Set 2, the bar and the checker-box move coherently to the same direction and at the same velocity (Figure 4-1b). In Set 3, only the checker-box moves while the bar remains stationary (Figure 4-1c).

We measured the contrast between the foreground and the background using the Michelson contrast function (Michelson 1995). We used a set of luminance values, with mean dark field luminance L_0, checker-box luminance L_{cb}, and bar luminance L_{bar}. The contrast of the bar is then calculated as:

$$C_{bar} = (L_{bar} - L_0)/L_0, \tag{4-1}$$

The contrast of the light background checker-box relative to luminance L_0 is calculated as:

$$C_{cb} = (L_{cb} - L_0)/L_0, \tag{4-2}$$

(Hupé et al. 1998) used a salience score is to indicate contrast difference between the bar and the checker-box, which is defined as:

$$Saliency = C_{bar}/C_{cb}, \tag{4-3}$$

To be consistent with the original study, we set $Saliency \in [1.0, 3.0]$ as low saliency, $Saliency \in [3.0, 4.5]$ as middle saliency, and $Saliency \in [4.5, 6.0]$ as high saliency. When $Saliency = 1.0$, the bar has the same intensity value with the background checker-box.

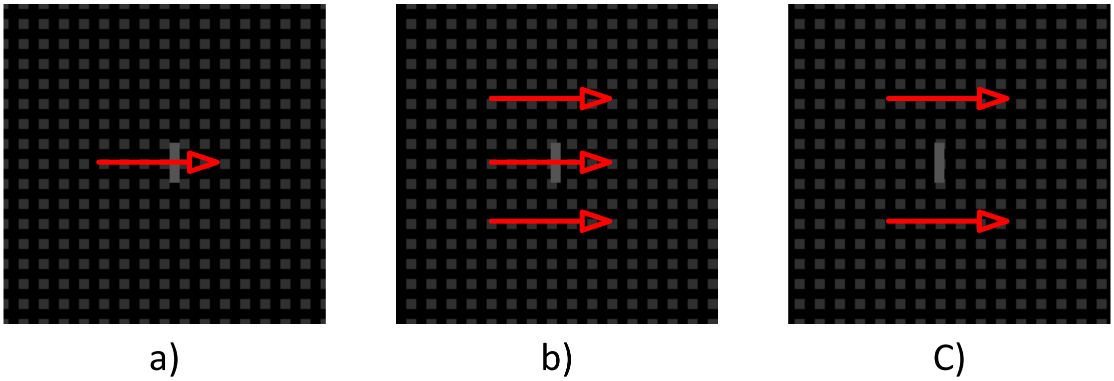


Figure 4-1 Test stimuli and three relative moving patterns. The stimuli include a solid bar that moves in front of a randomly distributed checker-box pattern at a constant velocity. a) Set 1: the bar moves on the stationary background. b) Set 2: the bar and the background move together. c) Set 3: the background moves and the bar remain stationary.

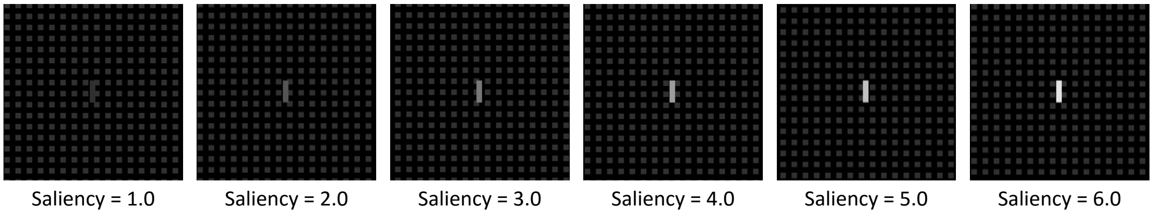


Figure 4-2 Stimuli at different saliency scores. At $Saliency = 1.0$, the bar has the same luminance value with the background checker-box. At $Saliency = 6.0$, luminance of the bar is approximate 4.75 times of intensity of the background checker-box.

When $Saliency = 6.0$, the intensity of bar is approximately 4.75 times of the intensity of the background checker-box. Figure 4-2 illustrates the relative intensity of bar and checker-box in different saliency settings.

In Hupé's experiment, orientation and size of stimuli are optimized to the receptive field of V1 neuron. In our case, we used a bar size of 8-pixel by 32-pixel, which are consistent with the modelled V1 receptive field. Stimuli move horizontally. The size of each checker-box is 8-pixel by 8-pixel. We ran the experiment 10 times for each setting. During each time, a background checker-box is randomly generated. Results reported in the following sections are the average value of the 10 runs.

4.1.2 Procedure

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. Table 4-1 summarizes the parameters of the proposed model. The modelled visual hierarchy includes the two visual pathways. The modelled dorsal pathway includes the retinal parasol ganglion cells (PGCs), the magnocellular cells of the LGN (M-cells), the dorsal layer of area V1 (noted as Dorsal V1), and area MT. The modelled ventral pathway includes the retinal midget ganglion cells (MGCs), the parvocellular cells of the LGN (P-cells), and the ventral layer of area V1 (noted as Ventral V1).

We did not include the other modelled visual areas discussed in Chapter 3 (i.e., area V2 and area MST). This is because the authors (Hupé et al. 1998) had attributed the observation to a mechanism between area MT and area V1. Thus, for the recurrence in our experiment, only that from the modelled area MT to the modelled area V1 was tested.

When choosing the recurrent pattern, we noticed that in the original experiment, the

Table 4-1 Experiment parameters used to configure the modelled neurons.

Modelled Visual Area	**Parameter Settings**
Parasol Ganglion Cells	2 spatiotemporal settings ($\sigma_s = 1.6$ and 3.2)
Magnocellular Cells	2 spatiotemporal settings ($\sigma_s = 3.2$ and 6.4)
Dorsal V1	12 spatiotemporal orientations with 30° apart.
Area MT	12 translation orientations with 30° apart 4 gradient orientations (0°, 90°, 180°, 270°)
Midget Ganglion Cells	2 spatiotemporal settings ($\sigma_s = 0.8$ and 1.6)
Parvocellular Cells	2 spatiotemporal settings ($\sigma_s = 1.6$ and 3.2)
Ventral V1	12 spatiotemporal orientations with 30° apart.

authors set both foreground and background stimuli to move to the same direction. This setting corresponds to the anisotropic modulation pattern (Pattern-1) of our proposal, where feedback is a single-direction representation.

In our simulation, we first investigated the output representations with and without early recurrence. Figure 4-3 compares model outputs for the three moving patterns, i.e., the bar moves in front of the stationary background, the background moves behind the stationary bar, and both the bar and the background move coherently. In the figure, a more reddish colour indicates a stronger neural response, while a more bluish colour indicates a weaker neural response. Consistent with the original experiment, we used a middle-to-high saliency score (e.g., $Saliency = 4.5$).

When the bar moves on the stationary background (Figure 4-3a), the non-modulated ventral V1 output representation (middle) contains a mixed energy distribution covering both stimuli and background regions. We noted that the bar region is more reddish than the background region. This is due that the ventral V1 responds actively to low temporal high spatial frequency visual information, thus its response to the bar is slightly stronger than its response to the background. In the modulated ventral V1 representation (right), neural response to the background is suppressed. However, since the dorsal representation has a coarse spatial resolution, a small portion of the background pixels surrounding the bar remains reddish. Our results show that the recurrent modulation indeed facilitates the ventral processing on the moving bar.

Figure 4-3b shows the case where the background check-box moves behind the bar while the bar is stationary. Compared with the Set 1 on the non-modulated ventral V1 representation (middle), neural response to the bar is weaker. The peak response to the bar is similar to the peak response to the background. In the modulated ventral V1 representation (right), due to that the dorsal pathway reacts to background motion, neural response to the background is enhanced. In addition, the boundary of the bar in both representations is broken. At first glance, this seems incorrect: as the bar is stationary, the dorsal pathway should not respond to the bar at all. However, this could be explained as the dorsal pathway detects the relative motion of the central bar with respect to the background. This is verifiable when we inspected the dorsal opponent energy representation, which clearly shows the counter-direction motion.

When the bar and the background move coherently (Figure 4-3c), responses to the background and the bar have been moderately extracted in the non-modulated ventral V1 representation (middle). In the modulated ventral V1 representation (right), response to the bar is enhanced: the bar region becomes more reddish. However, when compared with Set 1 (Figure 4-3a), the facilitatory effect from the dorsal representation is weaker.

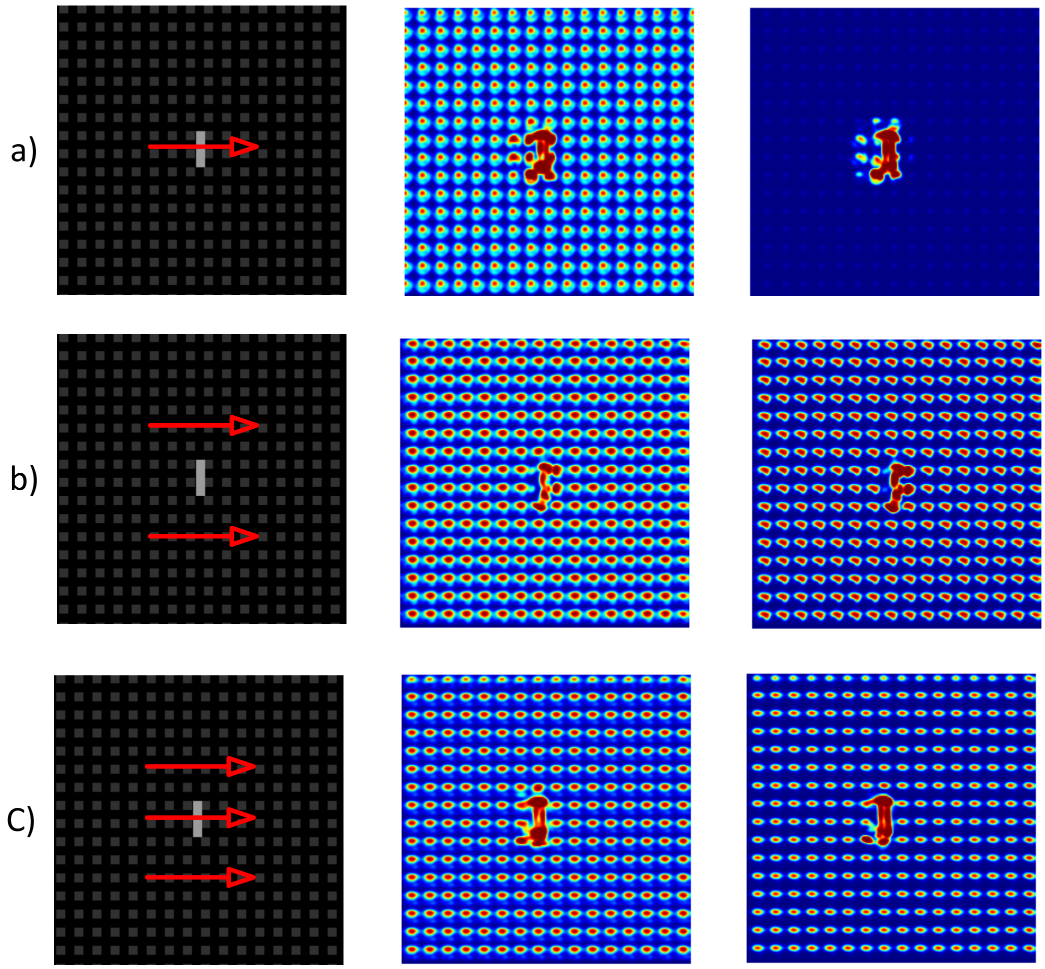


Figure 4-3 General observations of the three moving patterns. The left column shows stimuli as in Figure 4-1. These correspond to the three relative motion patterns. The middle and the right columns show ventral V1 response without and with early recurrent modulation, respectively. Reddish colours indicate stronger neural responses, while bluish colours indicate weaker neural responses.

4.1.3 Results

This section reports our quantitative analysis of model outputs at different saliency scores and at different moving velocities. In addition, we suspected that the proposed multiplicative inhibition may not be the only operation of early recurrence but might be the one that fits the current scenarios. Thus, we implemented an additive recurrent operation and compared it with the multiplicative operation.

Towards a fair comparison, responses of non-modulated and modulated ventral V1 representations are normalized (L1 norm). A centre region is defined that includes the bar and nearby background pixels. Output (pixel values) within the centre region and the whole region are accumulated respectively. We used the numerical ratio (rat_{er}) between the centre region strength and the whole region strength to indicate the selectivity of the target. Thus, $rat_{er} \in [0..1]$. In a straightforward manner, a higher ratio indicates that the response to the target is more significant. $rat_{er} = 1$ indicates that the response to the background is negligible. Alternatively, a lower rat_{er} indicates that the response to the bar is weaker relative to the surroundings.

We simulated the cell-recording experiment reported in (Hupé et al. 1998). We compared our results with theirs in a case-by-case manner, and concluded that the proposed model is capable of reproducing the original observations. In a sense, our simulation may be used as a piece of evidence to support the fast-brain hypothesis. Details are described as follows.

Bar with different saliency settings

When a salient bar (middle-to-high saliency score) moves on the stationary background, we could see that early recurrence has a facilitatory effect. We noted that the recurrence from area MT suppresses ventral V1 response to the checker-box, where rat_{er} is high. When the strength of recurrence is reduced, we observed decreased rat_{er} for the moving bar.

Compared to data reported in (Hupé et al. 1998), our results show a similar pattern of response suppression. An explanation of such an observation using the proposed model is that without early recurrence, information of motion and coarse spatial variation of the bar cannot reach the ventral V1. Consequently, the ventral V1 representation has less discriminative power to distinguish the bar from the background. As shown in the centre column of Figure 4-3a, the output ventral V1 representation includes responses to the background, thus the ratio rat_{er} is low. When the feedback is presented, the dorsal response is sent back to modulate the ventral V1. In this case, the ventral V1 response to the background stimuli is suppressed. As shown in the right column of Figure 4-3a, the modulated ventral V1 response is mostly focusing on the bar. In this case, rat_{er} is higher than that without early recurrence.

By visually inspecting the non-modulated and the modulated ventral V1 representations, we observed that the recurrent strength is dependent on the saliency score of the stimuli. Following the pattern of Figure 4-3a, Figure 4-4 shows that the suppressive effect is more significant in the low-saliency cases than in the high-saliency cases. At low-saliency, the centre bar is barely visible when both the bar and background move coherently. However, when the bar moves on the stationary background, the bar is clearly visible. We therefore extrapolated that area MT indeed provides information of motion and spatiotemporal variation, such that the visual system can use it to modulate ventral V1 processing.

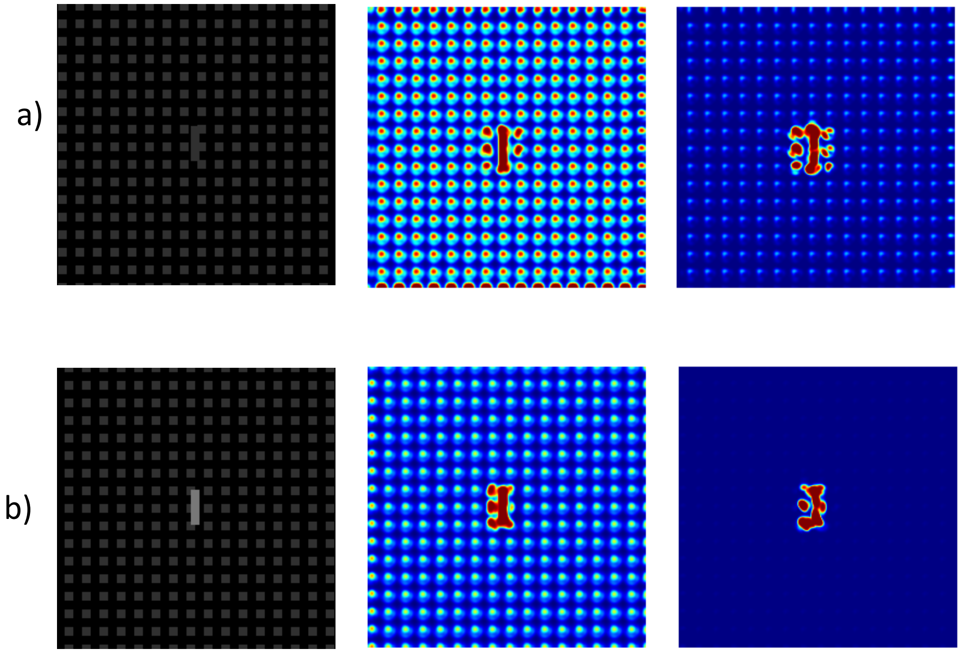


Figure 4-4 Saliency dependent suppression effects. a): Saliency = 1 and b) Saliency = 2. It shows that suppressive effect is more significant on low-saliency stimuli.

The original experiment showed a decrease of 39% in area V1 response when recurrence is absent. In our experiment, we observed similar decrease of rat_{er} in almost all saliency and speed settings. We computed rat_{er} for each saliency score and speed. For each setting, we repeated the experiment 10 times, each with a randomly generated background checker-box. rat_{er} values are average and reported in Table 4-2. We followed the original work to use the term "control" to refer to results with early recurrence and use the term "cool" to refer to results without early recurrence. From the table, we see that early recurrence boosts ventral V1 processing in all saliency and moving speed settings. In their original experiment, the authors only tested a single moving speed. In our case, we tested three moving speeds, where the bar moves at 1-pixel per sampled image is the optimized speed for the neuron that we constructed. When we increased the speed, we observed both ventral V1 responses with and without recurrence decreased as expected. However, the recurrent effect remains similar, except that in the 1-pixel setting, rat_{er} saturates to 1 when the saliency score is higher. Further, when the saliency score increases, we observed a decreased modulation effect. This is because at middle-to-high saliency scores, visual features extracted by the ventral pathway itself are already discriminative enough to segregate the bar from the stationary background. In all, the simulation indicates that early recurrence has the ability to improve ventral response in low saliency cases, consistent with what was reported by Hupé et al.

Table 4-2 Response changes from control (recurrence is present) to cooled dorsal areas (recurrence is NOT present). Scores are measured when the bar moves on the stationary background.

Speed / **Saliency**	**1-pixel per sampled image**			**3-pixel per sampled image**			**5-pixel per sampled image**		
	Control	Cool	Gain	Control	Cool	Gain	Control	Cool	Gain
1.0	0.879	0.309	64.8%	0.745	0.230	69.1%	0.715	0.213	70.2%
2.0	0.955	0.304	68.2%	0.856	0.268	68.7%	0.774	0.251	67.6%
3.0	0.987	0.435	55.9%	0.899	0.278	69.1%	0.796	0.265	66.7%
4.0	0.995	0.512	48.5%	0.929	0.291	68.7%	0.827	0.265	67.9%
5.0	0.998	0.624	39.0%	0.950	0.323	66.0%	0.846	0.292	65.5%
6.0	0.999	0.686	31.4%	0.967	0.354	63.3%	0.861	0.299	65.3%

In Hupé's experiment, the authors also observed modulatory effects in V2 and V3 a (Hupé et al. 1998). We suspected this may due to two reasons. First, it may be caused by direct recurrence from the dorsal pathway. Second, it may be caused by the modulation in area V1. Specifically, inconsistent visual features with the recurrent representation have been suppressed at area V1. Therefore, they will thus not reach area V2 and area V3, which leads to the observed modulatory effects.

Local context modulation effect with moving/stationary background

We knew that, due to network convergence, receptive fields of area MT are larger than those in the lower-level visual regions (e.g., area V1). Therefore, feedback from area MT facilitates ventral V1 processing with local context information. To test this effect, we compared the case where the bar moves on the stationary background (Set 1) with the case where the bar and the background move together (Set 2).

First, we noticed that the ventral V1 response to Set 2 was much weaker compared to Set 1. When the bar moves on the stationary background, the dorsal processing responds mainly to the bar and does not respond to the background. Via early recurrence, ventral V1 response towards the background is mostly suppressed. However, response towards the bar remains strong. In a test where the bar moves at 1-pixel per sampled image and the saliency score is 2, we observed an average response ratio $rat_{er} = 0.304$ when recurrence is absent, and $rat_{er} = 0.955$ when recurrence is present. This is a strong indication of an early recurrence effect.

However, when the stimuli and the background move coherently, the average response ratio is $rat_{er} = 0.159$ without recurrence, and $rat_{er} = 0.151$ with recurrence. Similar ratios are observed in all saliency settings. It seems in these cases, dorsal neurons respond to motion of both the bar and the background. Therefore, in the modulated ventral V1 representation, response to the surround pixels is not suppressed at all.

To further study this contextual modulation effect, we also investigated the case where the background moves and the bar remains stationary. Results show similar reversed modulation effect. The average response ratios are $rat_{er} = 0.139$ without modulation, and $rat_{er} = 0.127$ with modulation.

In (Hupé et al. 1998), the authors reported a similar push-pull recurrent effect. They noted that, as expected from the inhibitory interactions of many visual cortical neurons, the response to Set 2 was usually much weaker than when the bar moves on the stationary background. For a number of V3 neurons, inactivating area MT has a differential effect on the responses to these two stimuli. When the bar moves alone, the inactivation of area MT leads to response decrease. In contrast, response is enhanced when both the bar and the background move coherently. Further, in most cases, the modulatory effect when the background moves alone is null or very weak.

When studying early recurrence from a local context perspective, we observed that saliency is also an important factor in the suppression effect. When the early recurrent representation is inactivated, we noted a significant increase of response when the bar and the background move together. As we discussed, this effect is strong in low saliency cases but not in high saliency cases. Table 4-3 summarizes this pull-push effect when the bar and the background move together in varied saliency scores.

Table 4-3 Observed pull-push effect when the bar and the background move together with different saliency scores. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Speed / **Saliency**	**1-pixel per sampled image**		
	Control	Cool	Gain
1.0	0.131	0.134	-2.2%
2.0	0.151	0.159	-5.3%
3.0	0.217	0.209	3.7%
4.0	0.307	0.274	10.7%
5.0	0.405	0.341	15.8%
6.0	0.529	0.402	24.0%

Multiplicative recurrence vs. Additive recurrence

In Chapter 3, we hypothesized that for the type of early recurrent modulation, the essence of the operation is a multiplicative inhibition process that applies the fast-processed dorsal representation to suppress the slowly-processed ventral representation. An alternative way to model early recurrence is via an additive inhibition. Unlike the proposed multiplicative inhibition, an additive inhibition uses the dorsal representation as a gain factor over the ventral feed-forward representation. In this manner, the additive inhibition is defined as:

$$R'_{ventral} = \Theta[R_{ventral} + Inh_{dorsal}], \tag{4-4}$$

To test whether the additive inhibition has similar properties to the multiplicative inhibition, we repeated the experiment to compare results between the multiplicative inhibition and the additive inhibition.

When the bar moves on the stationary background, we observed comparable modulation effects in both additive and multiplicative inhibition. Table 4-4 - Table 4-6 compare the average response ratio at different speed settings. In these settings, both the multiplicative recurrence and the additive recurrence improve neural response to the moving bar. The multiplicative inhibition is slightly better than the additive inhibition.

When the bar and the background move together (Table 4-7 - Table 4-9), or the background moves but the bar is stationary (Table 4-10 - Table 4-12), we observed inconsistent results. We compared the average response ratio at different speed settings. We noted that the multiplicative inhibition yields different response ratios from the additive inhibition. When the bar becomes more salient, the multiplicative inhibition improves the bar responses. However, responses are reduced with the additive inhibition.

Table 4-4 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar moves on the stationary background at a speed of 1-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type	**Multiplicative Recurrence**			**Additive Recurrence**		
Saliency	Control	Cool	Gain	Control	Cool	Gain
1.0	0.879	0.309	64.8%	0.799	0.309	61.3%
2.0	0.955	0.304	68.2%	0.905	0.304	66.4%
3.0	0.987	0.435	55.9%	0.950	0.435	54.2%
4.0	0.995	0.512	48.5%	0.970	0.512	47.2%
5.0	0.998	0.624	39.0%	0.980	0.61	37.8%
6.0	0.999	0.686	31.4%	0.980	0.686	30.0%

Table 4-5 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar moves on the stationary background at a speed of 3-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type / Saliency	Multiplicative Recurrence			Additive Recurrence		
	Control	Cool	Gain	Control	Cool	Gain
1.0	0.745	0.230	69.1%	0.715	0.230	67.8%
2.0	0.856	0.268	68.7%	0.766	0.268	65.0%
3.0	0.899	0.278	69.1%	0.804	0.278	65.4%
4.0	0.929	0.291	68.7%	0.830	0.291	64.9%
5.0	0.950	0.323	66.0%	0.841	0.323	61.6%
6.0	0.967	0.354	63.3%	0.867	0.354	59.2%

Table 4-6 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar moves on the stationary background at a speed of 5-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type / Saliency	Multiplicative Recurrence			Additive Recurrence		
	Control	Cool	Gain	Control	Cool	Gain
1.0	0.715	0.213	70.2%	0.666	0.213	68.0%
2.0	0.774	0.251	67.6%	0.678	0.251	63.0%
3.0	0.796	0.265	66.7%	0.708	0.265	62.6%
4.0	0.827	0.265	67.9%	0.700	0.265	62.1%
5.0	0.846	0.292	65.5%	0.731	0.292	60.1%
6.0	0.861	0.299	65.3%	0.729	0.299	59.0%

Table 4-7 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar and the background move together at a speed of 1-pixel per frame. Additive recurrence test results are against our guess with reduced responses in higher saliency scores. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type / **Saliency**	**Multiplicative Recurrence**			**Additive Recurrence**		
	Control	Cool	Gain	Control	Cool	Gain
1.0	0.131	0.134	-2.3%	0.133	0.134	-0.8%
2.0	0.151	0.159	-5.3%	0.141	0.159	-12.8%
3.0	0.217	0.209	3.7%	0.161	0.209	-29.8%
4.0	0.307	0.274	10.7%	0.18	0.274	-52.2%
5.0	0.405	0.341	15.8%	0.198	0.341	-72.2%
6.0	0.529	0.402	24.0%	0.224	0.402	-79.5%

Table 4-8 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar and the background move together at a speed of 3-pixel per frame. Additive Recurrence test results are against our guess with reduced responses in higher saliency scores. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type	**Multiplicative Recurrence**			**Additive Recurrence**		
Saliency	Control	Cool	Gain	Control	Cool	Gain
1.0	0.129	0.135	-4.7%	0.132	0.135	-2.3%
2.0	0.179	0.164	8.4%	0.148	0.164	-10.8%
3.0	0.269	0.209	22.3%	0.167	0.209	-25.1%
4.0	0.353	0.253	28.3%	0.182	0.253	-39.0%
5.0	0.453	0.307	32.2%	0.199	0.307	-54.3%
6.0	0.556	0.35	37.1%	0.218	0.35	-60.6%

Table 4-9 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: the bar and the background move together at a speed of 5-pixel per frame. Additive Recurrence test results are against our guess with reduced responses in higher saliency scores. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type **Saliency**	**Multiplicative Recurrence**			**Additive Recurrence**		
	Control	Cool	Gain	Control	Cool	Gain
1.0	0.119	0.112	5.9%	0.132	0.112	15.2%
2.0	0.173	0.133	23.1%	0.149	0.133	10.7%
3.0	0.206	0.15	27.2%	0.161	0.15	6.8%
4.0	0.281	0.179	36.3%	0.177	0.179	-1.1%
5.0	0.361	0.218	39.6%	0.195	0.218	-11.8%
6.0	0.418	0.233	44.3%	0.208	0.233	-12.0%

Table 4-10 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: only the background moves at a speed of 1-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type	**Multiplicative Recurrence**			**Additive Recurrence**		
Saliency	Control	Cool	Gain	Control	Cool	Gain
1.0	0.117	0.128	-9.4%	0.122	0.128	-4.9%
2.0	0.127	0.139	-9.40%	0.123	0.139	-13.00%
3.0	0.14	0.145	-3.60%	0.131	0.145	-10.70%
4.0	0.165	0.162	1.80%	0.137	0.162	-18.20%
5.0	0.171	0.165	3.50%	0.137	0.165	-20.40%
6.0	0.195	0.181	7.20%	0.141	0.181	-28.40%

Table 4-11 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: only the background moves at a speed of 3-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type / **Saliency**	**Multiplicative Recurrence**			**Additive Recurrence**		
	Control	Cool	Gain	Control	Cool	Gain
1.0	0.133	0.139	-4.50%	0.132	0.139	-5.30%
2.0	0.163	0.147	9.80%	0.139	0.147	-5.80%
3.0	0.196	0.158	19.40%	0.146	0.158	-8.20%
4.0	0.257	0.174	32.30%	0.152	0.174	-14.50%
5.0	0.282	0.19	32.60%	0.157	0.19	-21.00%
6.0	0.317	0.2	36.90%	0.161	0.2	-24.20%

Table 4-12 A comparison between multiplicative recurrent modulation and additive recurrent modulation. Stimuli: only the background moves at a speed of 5-pixel per frame. Control indicates recurrence is present, and Cool indicates recurrence is NOT present.

Type	**Multiplicative Recurrence**			**Additive Recurrence**		
Saliency	Control	Cool	Gain	Control	Cool	Gain
1.0	0.147	0.125	15.00%	0.135	0.125	7.40%
2.0	0.19	0.146	23.20%	0.145	0.146	-0.70%
3.0	0.232	0.151	34.90%	0.152	0.151	0.70%
4.0	0.309	0.18	41.70%	0.153	0.18	-17.60%
5.0	0.328	0.186	43.30%	0.166	0.186	-12.00%
6.0	0.861	0.299	65.3%	0.729	0.299	59.0%

Among the centre-surround inhibition mechanisms discussed in the literature, addition is widely used to model how the surrounding response may inhibit the centre response, or vice versa. For example, the centre-surround LGN were modelled by using the Difference-of-Gaussians filter (Rodieck 1965, Davson 2012, Einevoll and Plesser 2012). The common cause to this type of inhibition is lateral mechanisms. However, the proposed model describes a different type of inhibitive mechanism, which utilizes feedback connections. This may be one of the reasons to explain why the additive inhibition is not as good as the multiplicative inhibition.

The proposed multiplicative inhibition also shares common characteristics to the gate control approaches of traditional top-down attention theories. For example, in VISIT (Ahmad 1991), the author proposed a number of gated features, which are computed based on a conjunction of lower-level features in constant time. Gated features are then used to inhibit other (unattended) features. In this manner, a gated feature map leads to a binary representation, which is used to multiply with up-coming features. As a result, irrelevant features are inhibited, leaving relevant features projects to higher cortical layers. The similar idea has been used in SERR (Humphreys and Muller 1993), where higher-level features are grouped to generate a "relative temperature map". In this representation, a "hot region" is more likely to become the next switch-on point and is therefore used as a gate or an anchor to group visual features.

4.1.4 Conclusion

Our results are consistent with the cell-recording experiments reported in (Hupé et al. 1998). This indicates that the proposed computational model of early recurrence is capable of explaining the feedback mechanism from dorsal area MT to ventral layers of area V1. Recurrence acts in a push-pull fashion, amplifying response to the optimal stimulus. The effect is stronger in low-saliency cases.

Our results support the hypothesis that feedback information from the dorsal areas facilitates the ventral processing of a bar moving on a stationary background. The boosting effect from the feedback is prominent, particularly in low-salience settings. In our view, the response of a higher-level ventral neuron is not simply determined by its feed-forward input or the local network via horizontal connections, but is also dependent on recurrence from the dorsal processing. Further, early recurrence in the form of multiplicative inhibition fits the neurobiology better than the additive inhibition. In conclusion, our simulation supports the Hypothesis 2 and the Hypothesis 3 proposed in Chapter 3: early recurrence is a multiplicative mechanism that applies information from the dorsal pathway to improve processing in the ventral pathway.

4.2 Kanisza Illusory Rectangles

In the literature, illusory contours (ICs) (Ginsburg 1975) have been suggested useful in studying how visual information is processed and integrated within the visual system. Using cell-recording techniques (Lee and Nguyen 2001, Ramsden et al. 2001) and neuroimaging techniques (Seghier and Vuilleumier 2006), the authors have identified a number of cerebral substrates that are relevant to the perception of ICs.

Based on dynamic property, ICs can be categorized into static ICs and dynamic ICs. A key challenge in studying static ICs is to localize the visual areas that cause the illusion. Most literature attributes this to the lower-level visual areas (Hirsch et al. 1995, Zeki 1996, Mendola et al. 1999). For example, in an experiment involving human subjects (Ffytche and Zeki 1996), the authors concluded that ICs are populated mostly in visual area V2. Essentially, the perception is a ventral processing mechanism.

Different from static ICs, dynamic ICs involve dorsal processing. Until the 1990s, our understanding of dynamic ICs was very limited. Although some studies attributed them with the dorsal pathway (Ramachandran et al. 1994, Goebel et al. 1998), many questions remain unclear, such as how the dorsal processing is involved and what their associated temporal relationships are.

Seghier and his colleagues made an important observation (Seghier et al. 2000b). They identified the visual areas that involve the perception of moving Kanisza rectangles using fMRI. The experiment included two sets of moving Pac-Man patterns, a stimulation pattern and a reference pattern, where the stimulation pattern causes illusion and the reference pattern does not. Their collected data suggested that the dorsal visual area MT and area MST involve a mechanism that causes ventral visual area V1 and area V2 to see the moving rectangle. They concluded that area V1 response is more prominent to animated illusory contours than to static illusory contours (Delon-Martin et al. 2000, Seghier et al. 2000a). Possibly, they hypothesized that this motion-cued illusion is originated from recurrence of the higher-level dorsal areas. This conclusion differed from a previous study from (Larsson et al. 1999), which suggested strong activation emerged from within area V1.

In this section, we report our effort to repeat the moving Kanisza rectangle experiment conducted in (Seghier et al. 2000b). Our intention is to understand whether the proposed computation can simulate the biological process that causes the illusion.

4.2.1 Stimuli

Visual stimuli include a set of 10 white Pac-Mans, placed in two vertical lines at a constant spatial distance. Each Pac-Man can rotate around its centre by 90° either in clockwise or counter-clockwise direction. In the stimulation pattern, four of the 10 Pac-Mans are arranged to create a static illusory Kanisza rectangle (Figure 4-5).

In a moving illusory Kanisza rectangle experiment, the 10 Pac-Mans rotate following two sets of patterns: a stimulation pattern and a reference pattern. In the stimulation pattern, at each sample image, all Pac-Mans in the left column rotate clockwise by 90°, and all Pac-Mans in the right column rotate counter-clockwise by 90°. In this way, the illusory Kanisza rectangle moves downwards (Figure 4-6).

In the reference pattern, each Pac-Man is tilted by 45° (Figure 4-7). Otherwise, Pac-Mans move in the same way as in the simulation pattern. This set-up ensures that in all sampled images, the 10 Pac-Mans never show an illusory pattern: the stimuli disallow the perception of the illusory Kanisza rectangle.

Note that in the reference pattern, there would be no chance for any illusory Kanisza rectangle to present. However, we are interested in investigating the model's behaviour when a single illusory Kanisza rectangle presents at one sampled image but disappears in the next. Therefore, a third transient pattern is introduced. In this pattern, each Pac-Man rotates 45 degrees for each sampled image (Figure 4-8).

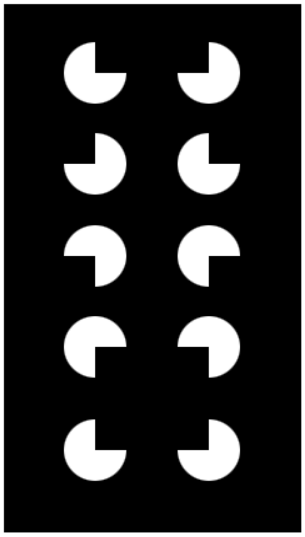


Figure 4-5 An example of static illusory Kanisza rectangle. Experiments in the literature attributed the perception of the dark rectangle enclosed by the four Pac-Mans as a mechanism in the ventral pathway.

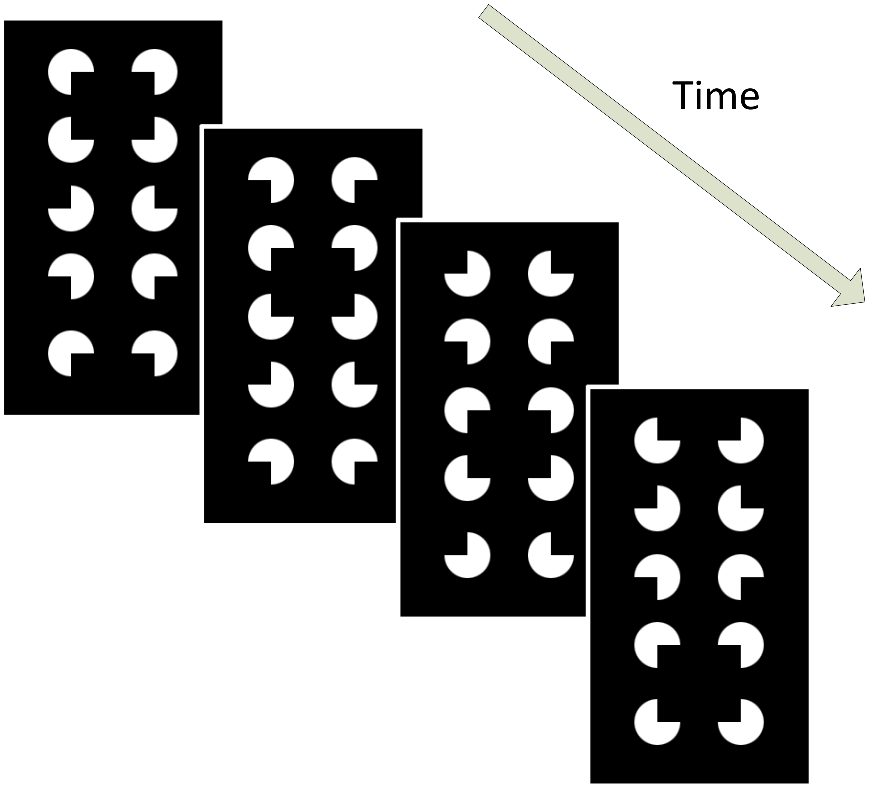


Figure 4-6 An example of moving illusory Kanisza rectangle. During the experiment, Pac-Mans rotate in a stimulation pattern, such that the dark rectangle moves downwards.

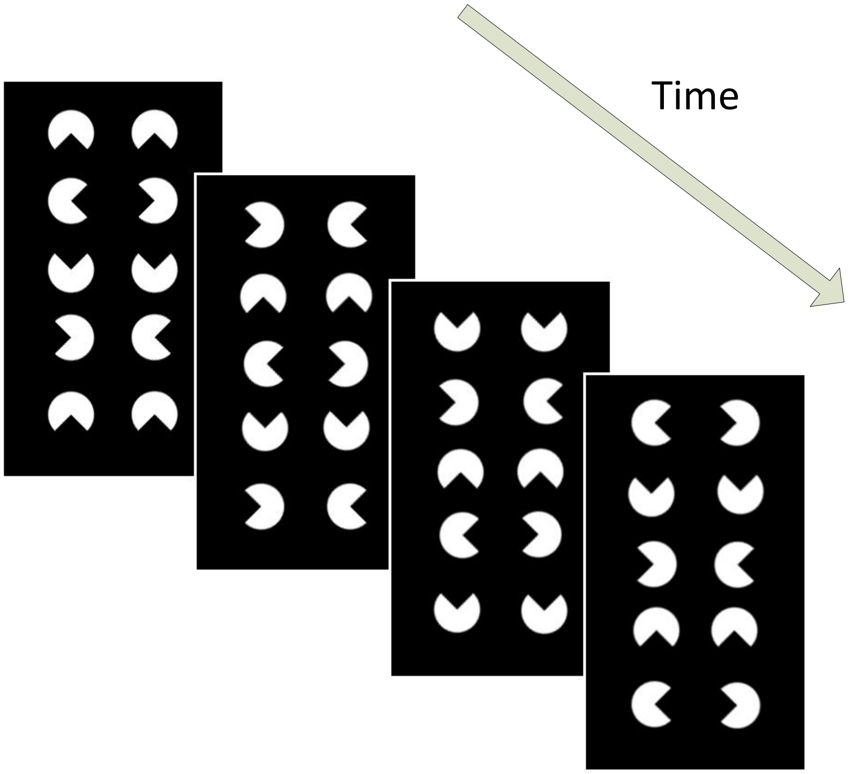


Figure 4-7 An example of moving Pac-Mans in the reference pattern. The Pac-Mans are tilted by 45 degrees. This setup ensures that throughout the experiment, no illusory Kanisza rectangle will be perceived.

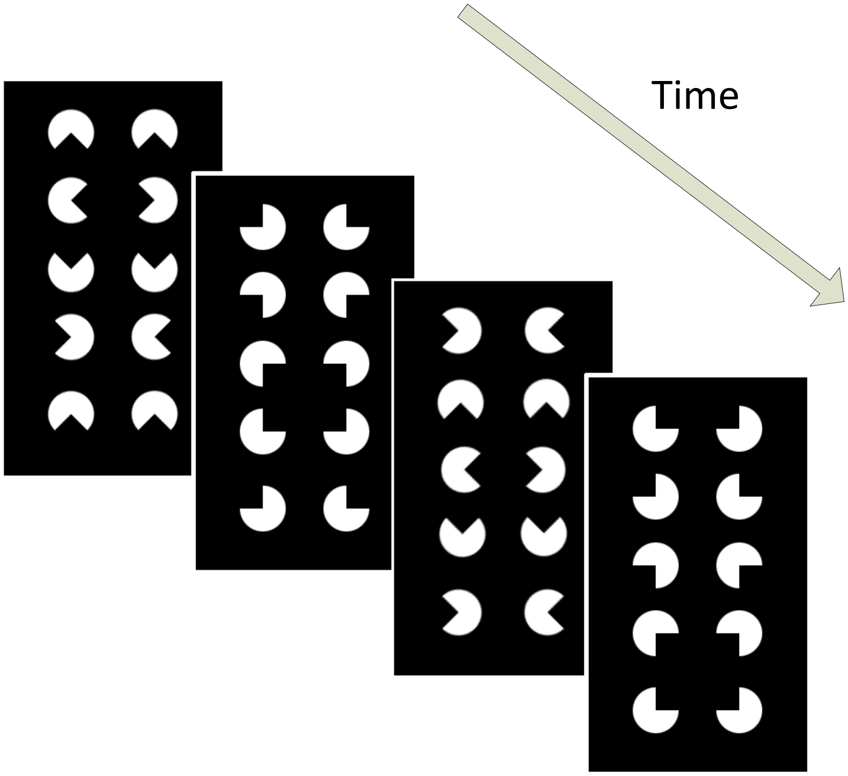


Figure 4-8 A third rotation pattern, the transient pattern. Illusory Kanisza rectangle shows at one stage of the refresh and disappears in the next.

4.2.2 Procedure

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. In (Seghier et al. 2000a), from the subject's view, the illusory rectangle extends 2° horizontally by 1.7° vertically, and the rectangles were displayed symmetrically with respect to the vertical meridian of the two columns. The radius of each Pac-Man extends approximately 0.45°. Thus, the support ratio of contour (Shipley and Kellman 1992), which is the ratio of the total length of the borders, actually show to the perimeter of the rectangle, is 0.48. To make our experiment consistent, we used a similar ratio. Each Pac-Man has a diameter of 50 pixels. The distance between horizontally adjacent Pac-Mans (measured between the two Pac-Mans' centre points) is 90 pixels, and the distance between vertically adjacent Pac-Mans is 75 pixels.

4.2.3 Results

In Seghier et al experiment, the authors identified two distinctive regions that may cause the illusory perception: visual areas V1/V2 and visual areas located far in the dorsal pathway (corresponding to visual area MT/MST). They noted that the stimulation pattern leads to strong activation to the illusory contour. Further, response in lower-level visual areas to the moving IC is stronger than that to the static IC. In our experiment, we have observed comparable results.

In the static Kanisza illusory experiment, the dorsal pathway was "deactivated" due to lacks of motion information. Visual features of the Pac-Mans are solely computed in the ventral pathway. However, since the receptive field of area V1/V2 is relatively smaller than that of the higher-level ventral areas, it is unlikely that these regions perceive the illusory rectangle. As shown in Figure 4-9, output representation of area V1/V2 extracts the fine details of Pac-Mans. Our modelled V1 representation provides a collection of edges that construct Pac-Mans. The modelled V2 representation further computes the curvature information and end-stopped information, such that it is capable of showing smoothed curves and highlights corners. However, both representations fail to echo the IC. We thus conclude that the perception of the illusory contour is unlikely to emerge by area V1 or area V2. Instead, it may be the result of higher-level ventral processing.

In the moving Kanisza illusory experiment, the dorsal pathway responds to motion and sends results to modulate lower-level ventral areas. By the recurrent modulation, we observed strong activations in area V1 and area V2. However, the strong activations only show in the stimulation pattern, not in the reference pattern. Interestingly, the response to the transient pattern is weak, even if there is no illusory rectangle.

Figure 4-10 illustrates the response of dorsal V1 neurons. Note that the dorsal V1 has a coarse spatial accuracy: the region lit around each Pac-Man is blurry. Although it reacts to motion, its output representation does not contain the moving illusory contour. This is because the receptive fields at this level are very small.

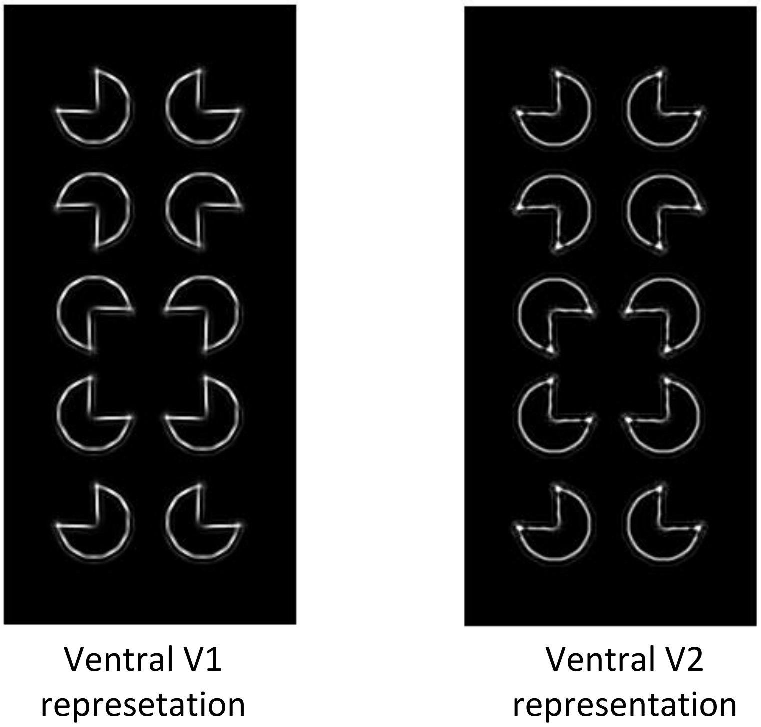
Along the dorsal pathway, area MT integrates motion information from area V1 (Figure 4-11). Although an MT neuron has a larger receptive field, it still cannot cover the complete illusory region. In contrast, area MST has a clear response to the illusory rectangle (Figure 4-12). However, at this level, spatial information is lost. From the brightness, it is hard to localize or recognize the actual contour of the Pac-Mans.

Figure 4-13 illustrates the early recurrent modulation between area MST and area V2. Immediately, since area MST has strong response to the illusory contour, in the modulated

V2 representation, curvatures of Pac-Mans defining the rectangle are much more brightened. We compared the modulated and the non-modulated V2 representations over the corners of each Pac-Man. In absence of modulation, the corner pixels of each Pac-Man are brighter. This is due to the V2 end-stopped cells responding strongly to the ending point of an edge. However, since MST neurons respond actively to the illusory region, recurrence from area MST plays a role that inhibits V2 to ending points within the region. We further hypothesized that a lack of ending points may facilitate higher-level ventral computation to close the missing contour.

In the reference pattern experiment (Figure 4-7), no illusory rectangle presents at any sampled images. We observed no significant activations in area V2 during the whole process (Figure 4-14). Although the dorsal regions respond to motion, their responses remain localized. In the output representation of area MST, it is difficult to distinguish any region based on motion cues. In this case, the feedback representation has very limited impact on the ventral processing, which is consistent with the results reported in (Ffytche and Zeki 1996, Seghier et al. 2000a).

However, with the transient rotation pattern (Figure 4-8), we observed different results. As shown in Figure 4-15, the MST neurons respond to motion and perceive the illusory contour. However, activation to the illusory contour is weaker compared with that in the stimulation pattern. This may due to the distracting samples between the two illusory samples. The result of area MST feeds back to modulate area V2 processing. Note that in Figure 4-15, edges contributing to the illusory rectangle are brighter. Further, in the next sampled image where no illusory rectangle presents, the very same region remains active.



Ventral V1 represetation

Ventral V2 representation

Figure 4-9 Ventral pathway output representations to static Pac-Man input.

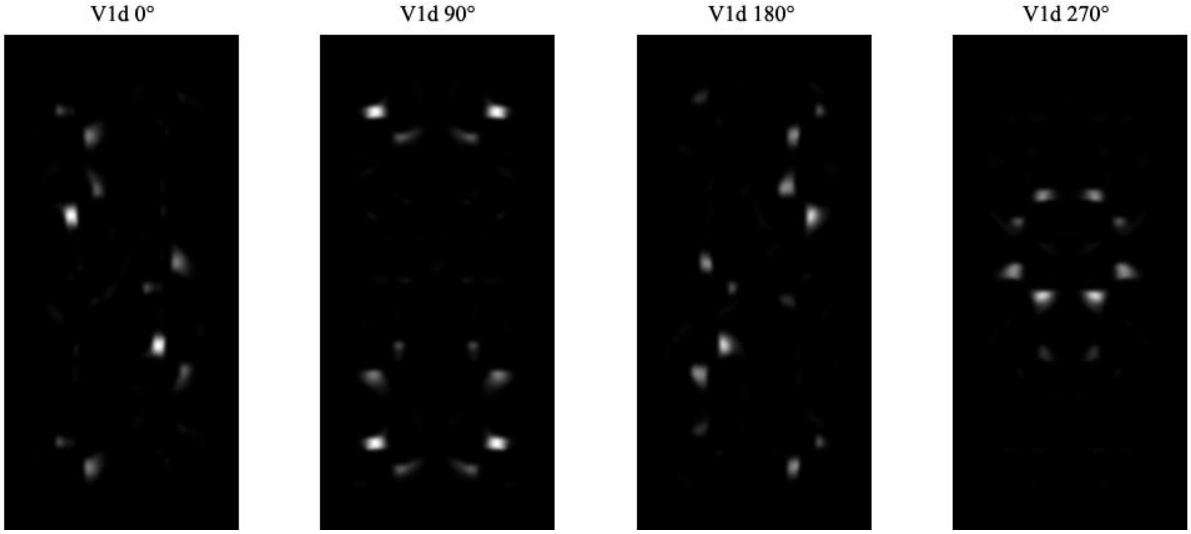


Figure 4-10 Dorsal layers of area V1 responses to the moving Pac-Mans for the stimulation pattern. Brighter areas indicate neural responses to motion toward the four directions.

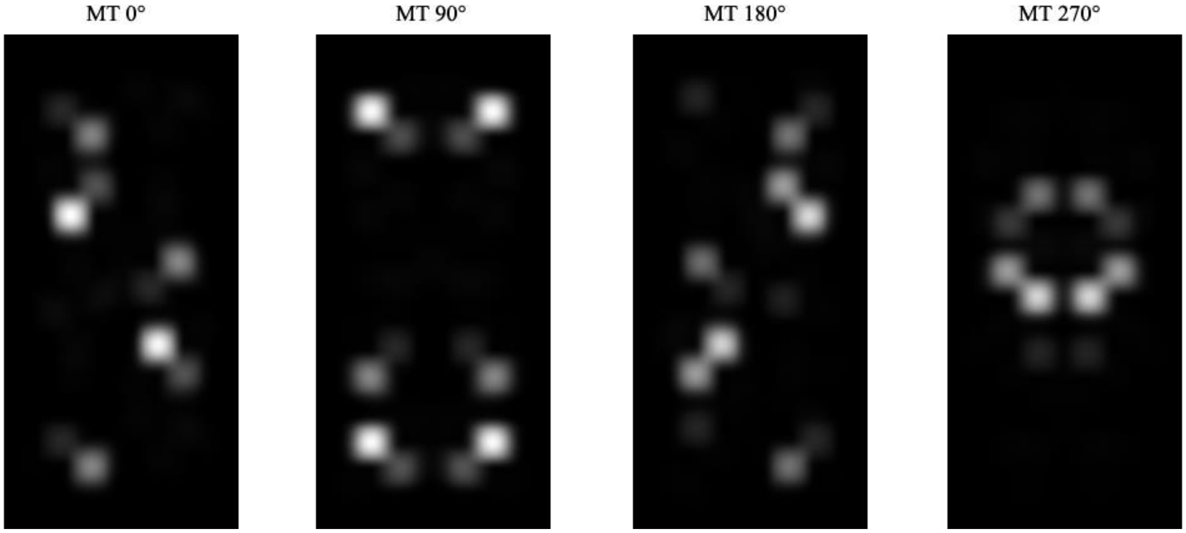


Figure 4-11 Area MT neural responses to the moving Pac-Mans for the stimulation pattern. Brighter areas indicate neural responses to motion toward the four directions.

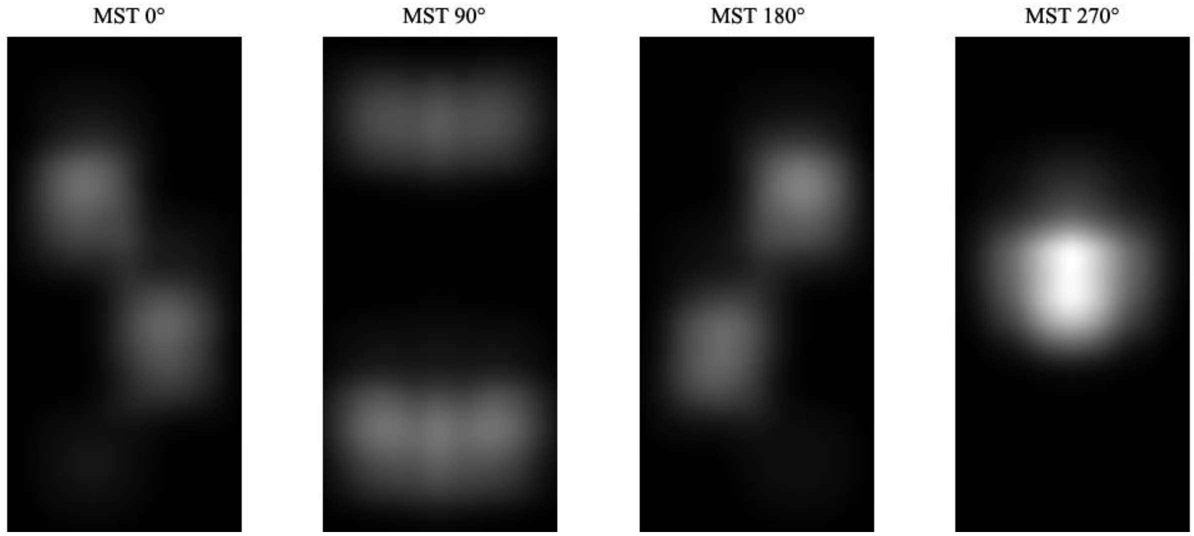


Figure 4-12 Area MST neural responses to the moving Pac-Mans for the simulation pattern. Note there is a strong activation over the illusory contour region.

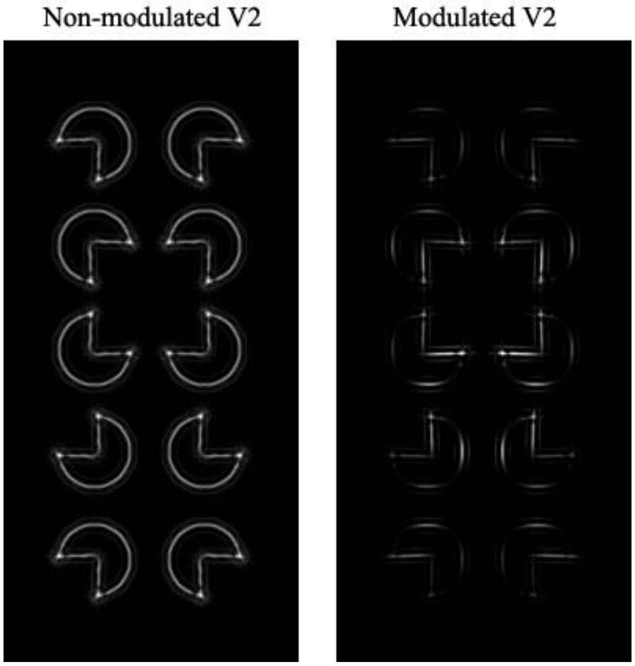


Figure 4-13 Non-modulated (left) and modulated (right) V2 representations for stimulation pattern. It clearly shows that early recurrence causes an activation over the illusory contour region in the modulated V2 representation. The result shown here is in agreement with (Seghier et al. 2000a).

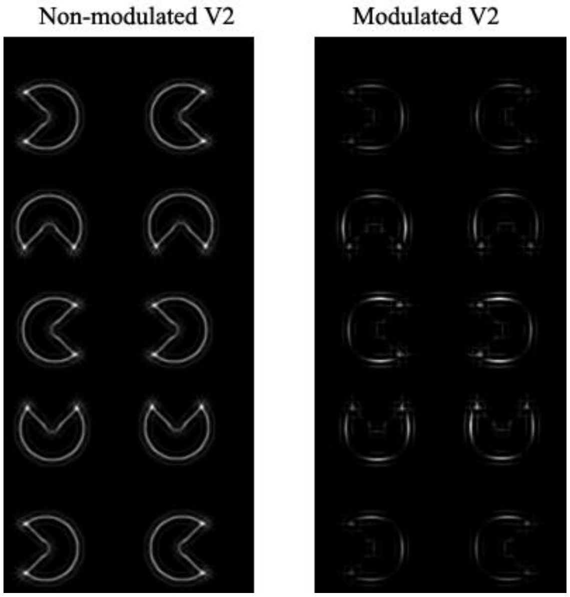


Figure 4-14 Non-modulated (left) and modulated (right) V2 representations for with reference pattern. Since there is no illusory contour throughout the simulation, there is no significant activation in the modulated V2 representation.

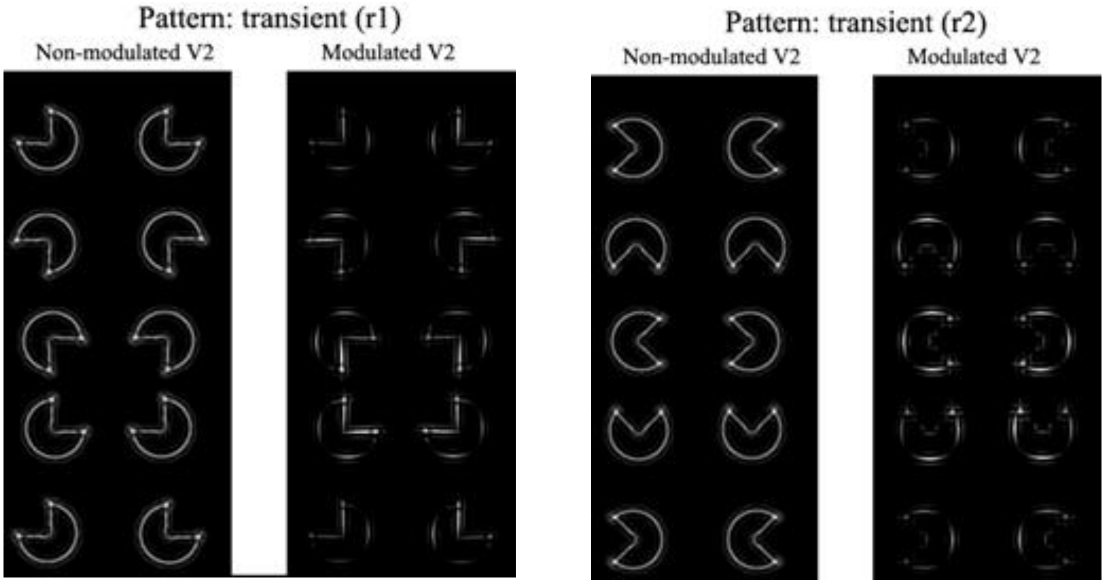


Figure 4-15 Non-modulated and modulated V2 representations for the transient pattern. We show two refreshments: (r1) includes an illusory contour, and (r2) is without the illusory contour. In both cases, the contour regions have been activated, indicating that early recurrence may cause illusions even without illusion presence.

4.2.4 Conclusion

To answer the question of what causes the perception of the moving illusory contour, it is important to analyze the phenomenon from a spatiotemporal perspective. On one hand, the spatial extent of the receptive field is important. It leads to the region in the visual system that is more likely to cause the illusion. On the other hand, the temporal correlation across multiple visual areas may provide insight into the temporal delays between the two visual pathways.

In this regard, existing studies of illusory contours can be divided into two categories: those supporting the hypothesis that illusory contour perception is caused by localized-and-lower-level mechanisms (Lesher and Mingolla 1993, Matthews and Welch 1997, Rajimehr 2004), and those supporting global-and-higher-level mechanisms (Grill-Spector et al. 2001, Vuilleumier et al. 2001, Han et al. 2002).

For the localized-and-lower-level mechanism, it is assumed that only the early visual areas (such as area V1 and area V2) are involved. Evidence from functional imaging and cell-recording supports this mechanism (Seghier and Vuilleumier 2006). The cause of illusory contour perception is that the receptive fields of lower-level areas are small and retinotopically organized. Their functional properties may afford an efficient detection of local details. The local details include edge and contour, which provide a critical input for the generation of illusory contours. Further, because of this localized property, lateral information is very limited. The initial representation of edge and contour may thus be associated with illusory information, based on relative brightness and contrast information.

The global-and-higher-level mechanism involves higher-level visual areas. Vision generated at higher-level visual areas covers a larger receptive field. At the top level, representation of global scene structure becomes available. The higher-level representation may also involve experience, task or higher-level visual mechanisms. For example, a neuroimaging experiment (Grill-Spector et al. 2001) shows that the lateral occipital complex is a visual area that generates localized illusory contour representation based on its shape recognition results. Global information feeds back to the lower-level areas in V1/V2 to influence the figure-ground segregation. This reinjection of global information into lower-level analysis may be important for the brain to see an illusory contour.

Here, we believe that the temporal delay is an important factor to separate those two mechanisms. For the first mechanism, the initial representations of area V1 and area V2 are generated within the first 100 milliseconds after the stimuli onset (Lee and Nguyen 2001, Foxe and Simpson 2002). For the second mechanism, since the feedback comes from higher-level visual regions, the perception of illusory contour should require a much longer time (Murray et al. 2002, Pegna et al. 2002). Based on this clue, in the current study, we propose that for moving Kanisza illusory rectangles, the cause is neither localized-and-lower-level mechanisms nor global-and-higher-level mechanisms. Instead, early recurrence is likely to be the real cause.

4.3 Discussion

The proposed recurrent computation manifests itself as a lower-level feedback process during a feed-forward sweep of visual information processing. The temporal delays between the dorsal and the ventral areas, and the availability of cross-pathway recurrent connections allow this early recurrent mechanism to be at play. In this work, we hypothesize that the recurrent operation between the dorsal area and the ventral area takes a form as multiplicative inhibition. Early recurrence suppresses ventral processing. As such, only consistent visual features remain in higher-level ventral representations.

At first glance, the proposed model shares many common characteristics with Hubel and Wiesel’s visual hierarchy. However, their model describes a feed-forward visual processing paradigm without considering the temporal latency between the two pathways. As more evidence supports the temporal dynamics during early visual processing and supports the diversified feature processing between the visual pathways, we investigated pathway interactions during a feed-forward swipe. A key motivation to the current work is to formalize how the visual hierarchy utilizes shortcuts between the dorsal pathway and the ventral pathway to facilitate ventral processing. In this sense, the current work does not stand opposed to Hubel and Wiesel’s idea. Rather, we have proposed a recurrent mechanism that takes place during the feed-forward processing.

Early recurrence provides a mechanism to refine the ventral representation. We noticed that the refinement may be achieved by feed-forward selective mechanisms, such as H-max (Riesenhuber and Poggio 1999, Serre et al. 2007) and signal pooling (Barlow and Tripathy 1997). Although the goal of H-max and signal pooling is to some degree similar to the proposed model, there is no concept of early recurrence in those models. In those works, winning features are selected based on intrinsic feed-forward analysis. In our view, early recurrence proceeds in a different way. The selectivity is achieved by surround suppression wherein the suppression comes from the dorsal pathway. Although the recurrent representation is based on feed-forward information, features computed in the dorsal and the ventral pathways are of different natures. For example, we have shown how to use motion information to facilitate object processing.

Alternatively, lower-level feature representation refinement may be performed via lateral suppression. The current work is not opposed to such idea. Early recurrence provides an additional approach to improve early visual representation. In future work, we may further investigate the property of early recurrence to distinguish it from lateral suppression.

From a context facilitation perspective, early recurrence provides a mechanism to apply localized dorsal information to facilitate ventral processing. Based on the physiological properties of the dorsal pathway, the dorsal representation is a description of input: it catches the perceptually salient information but lacks spatial accuracy. On the other hand, the lower-level ventral areas compute more detailed spatial variations, which are then constructed into edges, corners, curvatures, and shapes. Without a selective mechanism, information of target and background are mixed. When early recurrence presents, we propose that the dorsal representation is used to suppress the ventral processing of background features: ventral neurons correspondent to spatial locations not in favour of dorsal processing (non-salient regions) are inhibited. This modulation takes place at the very low level of the visual hierarchy, and thus the refinement is of the simplest form of visual features. In our experiment, we showed how edge representation is refined such that only target-related edges remain in the modulated representation.

Following the proposed model, we further investigated its utility in computer vision. The scope of computer vision applications that may utilize the proposed mechanism is very broad. Most feature-based vision systems may take advantage of the proposed operation to improve their representations. Multiplicative inhibition is simple in definition and has low computational complexity.

We will present two computer vision applications in the following chapters. These applications utilize early recurrence to refine simple visual features, which are then used in higher-level modules for different purposes. The intention here is to demonstrate that 1) early recurrence is a promising technique to boost performance, and 2) the proposed operation is easy to implement to fit these applications.

Chapter 5. Impact of Early Recurrence on Visual Computation

In the previous chapters, we have proposed the model of early recurrence with its associated computational components. By simulating two biological experiments, we showed that the proposed computation is consistent with biology.

From this chapter, we report our efforts to apply the computation to improve computer vision systems. This chapter focuses on the topic of visual saliency and two related applications: background subtraction and scene recognition.

5.1 Early Recurrence Improved Visual Saliency Representation

Visual saliency is a subjective perceptual quality measurement. It is a representation to indicate how visual stimuli appear different from the rest of the visual field (Itti 2007). A classic usage of visual saliency is to detect region-of-interest. Detected (salient) regions facilitate the system with reduced search space and computational burden, which may ultimately improve the overall system performance. In this context, an optimal saliency representation picks all and only the content of interest. A sub-optimal saliency representation, on the other hand, may contain either false positives (regions that do not contain content of interest) or false negatives (content of interest that is not picked at all).

In this chapter, we first show that early recurrence can be used to improve different types of visual saliency representations. We then use background subtraction and scene recognition as examples to further show how the visual saliency representations can be used to boost practical system performance.

In the literature of neuroscience, most studies attribute visual saliency as a lower-level representation based on simple features (Li 2002, Treue 2003). One of the theories holds that visual saliency is computed within the primary visual cortex (Li and Snowden 2006). Further, since visual saliency is closely related with object perception, it is commonly deemed as a mechanism in the ventral pathway (Treue 2003, Vanrullen 2003).

Visual saliency is also related to the concept of visual attention (Koch and Ullman 1985). In a number of saliency-based attention models (Itti et al. 1998, Zhang et al. 2008, Bruce and Tsotsos 2009a), it is the driving factor that cues region-of-interest selection.

In an early psychological study, visual saliency is described as a master map representation (Treisman and Gelade 1980b). The idea was originally proposed to determine attended locations. The theory states that visual information extracted via different visual pathways forms a centralized map. The intensity values on the map represent how important or salient a region is compared with its surrounding regions. Based on this master map representation, the region with the strongest intensity value is selected as the attended location. The process is analogous to a spotlight mechanism. This theory motivated a computational representation (Koch and Ullman 1985). In addition to the saliency map calculation, the model also includes winner-take-all algorithm to pick the most salient region (the winner) as the attended location.

A further development (Itti et al. 1998) applies saliency information to predict eye fixations. The core computation includes a feature extraction module and a winner-take-all module. Different filter-based techniques extract visual features such as intensity, colour, motion, and shape. Each extracted feature forms a feature map representation: an intensity image measuring the feature distribution. A saliency map is then computed based on weighted summation over all feature maps (Itti and Koch 2001). The winner-take-all module then selects the most prominent set of pixels as the attended region. Based on an experiment on human subjects, the authors concluded that the proposed computation is capable of predicting eye fixations. Bruce and Tsotsos studied visual saliency from an information theory perspective (Bruce and Tsotsos 2009a). In their model, visual saliency is measured as center-surround entropy, or self-information of a given image. Further, the authors proposed to construct feature maps using filters learned via independent component analysis (ICA).

5.1.1 Rationale

In Chapter 3, we proposed the core concept of early recurrence: to apply results of dorsal computation to refine early ventral processing. For the application of visual saliency, we hypothesize that the dorsal processing (on motion and coarse scale-spatial variation) has strong impacts on the saliency computation.

Of course, we are not the first to include motion in the saliency representation. In (Koch and Ullman 1985) and its computational implementation (Itti et al. 1998), motion has already been used. However, the way they use motion is debatable.

In Itti et al., the saliency map is constructed by combining all feature maps. During this process, motion is just one of the many features the model considers. The whole computation manifests itself as a feed-forward mechanism. Motion does not have the priority in constructing the saliency map; and it is not used to modulate other feature representations. This simple approach seems to be effective. In a test case where an object has a simple motion pattern in front of stationary background, the target location will be easily picked up as the fixation. However, the performance will decrease if input becomes complicated. The saliency map will be heavily “polluted” by the intricate and scattered motion patterns. To solve this problem, there have been many attempts to improve the quality of the feature maps. For example, one can use more complicated motion filters, or even rely on learning algorithms, to correlate with the motion. However, this approach leads to increased computational costs, and does not fundamentally correct the behaviour.

During our investigation, Bullier’s fast-brain hypothesis motivated us (Bullier 2001). The notions of fast dorsal processing and lower-level recurrent connectivity inspired us with an alternative approach: unlike Itti et al., we use motion and coarse-scale spatial variation as feedback representation to modulate the saliency computation.

By comparing output of the modelled dorsal pathway with human-labelled ground truth, we observed that the dorsal representation has a marked correspondence with visual saliency. However, the representation lacks the spatial accuracy. At the same time, output of the modelled ventral pathway contains background, cluttered scene-parts, and image noise in details. By early recurrence, a substantial number of the non-salient pixels can be suppressed, leaving details of the salient target untouched. That is, via early recurrence, a spatially-accurate yet perceptually-salient representation is computable.

To investigate whether the proposed early recurrent operation fits the role, we applied our computations to several existing visual saliency models. We used real images and surveillance videos to verify our model. We used standard matrices to evaluate saliency performance. To draw a quantitative conclusion, we include human fixation data to compute correlation between model output and biological observations.

5.1.2 Experiment

Following Chapter 3, we implemented a number of visual areas within the two main visual pathways. Implementation of the dorsal pathway includes the magnocellular layer of the LGN, the dorsal layers of area V1, and the middle temporal cortex (MT). Implementation of the ventral pathway includes the parvocellular layers of the LGN and the ventral layers of area V1.

The recurrent computation further relies on the facts that: 1) the two visual pathways compute different visual features, 2) the pathways have different conduction speeds, with the dorsal pathway conducting signals faster than the ventral pathway, and 3) the network allows to use the dorsal representation to modulate the ventral processing.

Figure 5-1 illustrates the visual hierarchy of the implemented model. We calculated two sets of visual features separately, corresponding to the dorsal features and the ventral features. They are defined using different spatiotemporal scales. Specifically, to be consistent with the neuroscience literature (Derrington et al. 1984), the dorsal pathway computes high-temporal-low-spatial frequency scale features (i.e., coarse spatial variation and motion) and the ventral pathway computes low-temporal-high-spatial frequency scale features (i.e., fine pixel variation). Details of the computation are in Chapter 3.2.

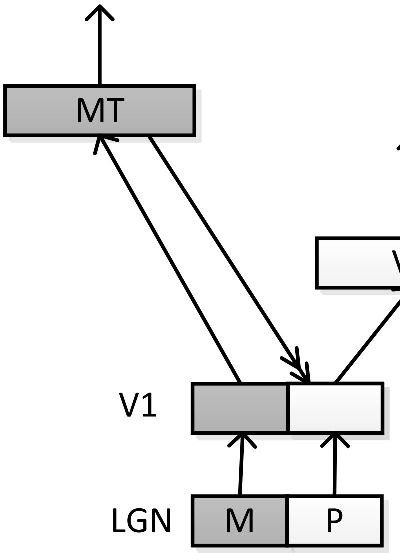


Figure 5-1 Connections between dorsal and ventral pathways considered in this work. Grey blocks denote dorsal areas, and white blocks denote ventral areas. Lines denote connections, particularly the double arrow line denotes early recurrence from area MT to area V1.

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. The purpose of the evaluation is to determine whether early recurrence is generally applicable to boost existing saliency models. In our evaluation, we used three existing models (Itti et al. 1998, Zhang et al. 2008, Bruce and Tsotsos 2009a), which compute saliency from different perspectives. In particular, saliency in (Itti et al. 1998) is defined as strength of summed visual feature activations, while in the other two proposals (Zhang et al. 2008, Bruce and Tsotsos 2009a), visual saliency arises from measuring the self-information (but in different manners) based on natural image statistics. In our comparison, saliency maps calculated by these original models provide us with a baseline for performance. It is important to note that the ventral computation concerned in the current model represents the conventional kind of feature processing seen in the three existing models. In our proposal, it is the modulated ventral representations, rather than their original versions, that generate the final saliency map.

The proposed computation fits itself easily into the existing models by applying the recurrent representation to modulate feature processing. This is such that in the revised models, saliency representations are calculated based on the modulated feature maps. One can then evaluate to what extent early recurrence improves saliency performance over baseline scores. The implementations are tested with a set of cluttered images that were introduced in (Bruce and Tsotsos 2009a) to evaluate static stimuli and are also tested with a set of surveillance videos from YouTube to evaluate spatiotemporal stimuli.

We measured saliency performance by the receiver operating characteristic (ROC) curve, which have been widely used in related work. For a given ground truth (G) and a normalized saliency map (S), by varying the threshold $\delta \in [0..1]$, a smooth curve is generated as true positive rate $G(x, y) \geq \delta$ and $S(x, y) \geq \delta$ versus false positive rate $G(x, y) \leq \delta$ and $S(x, y) \geq \delta$, where (x, y) gives a pixel's coordinates.

Figure 5-2 shows the improved saliency maps given static input images. Saliency maps produced by the original models and the modulated versions are paired in groups. Reddish pixels indicate salient regions. We see that as visual input becomes complicated, saliency calculated by the baseline models become less discriminative and less correlated with fixation density maps recorded on human subjects (Bruce and Tsotsos 2009a).

We also see that the modulation leads to more similarities between object regions and reddish regions in the modulated saliency maps (right image of each pair) than those in the original saliency maps (left image of each pair). The main difference between a pair of saliency maps is that a substantial number of background pixels are suppressed in the modulated saliency maps. Although salient regions calculated by the three baseline models are different, in the modulated versions, salient regions are all confined to the recurrent representations, leaving the remaining regions mostly in blue (not salient).

It is interesting to note that although early recurrence leads to improved saliency representations, the recurrent representations themselves are not always matching with the eye fixations. In our examples, the recurrent representation usually yields a coarse spatial scale description. This is consistent with our understanding that the dorsal representation is indeed lack of visual accuracy. However, the dorsal representation provides the desired context of surroundings. In a way, this context facilitates the ventral processing to respond to salient targets.

Mean ROC curves have been generated based on human fixation densities (Bruce and Tsotsos 2009a). From the ROC plots of Figure 5-2, it is obvious that curves produced by the modulated saliency maps (solid lines) augment their original versions (dashed lines) significantly. Areas under the curves are calculated. We see that the modulation increases the area for all three methods, which confirms that early recurrence is effective in improving these saliency measurements.

Figure 5-3 illustrates improved saliency calculation for spatiotemporal inputs (videos). Test samples include videos from various viewing angles and under different illuminant conditions. Targets (i.e., vehicles and pedestrians) are manually labelled as ground truth for evaluation. We see that the recurrent representations in each test clearly highlight regions of moving stimuli, leaving stationary and cluttered scene parts suppressed. The improvement over the original saliency model is obvious when we compare the mean ROC curves (red lines versus blue lines). In this investigation, we also compared a reference model similar to (Bruce and Tsotsos 2009b), where saliency is computed using output of spatiotemporal filters (green lines).

In conclusion, we conducted empirical and quantitative comparisons to study the facilitatory effect of early recurrence in saliency calculation. From the modulated saliency maps and the associated ROC curves, we concluded that early recurrence leads to significant improvements.

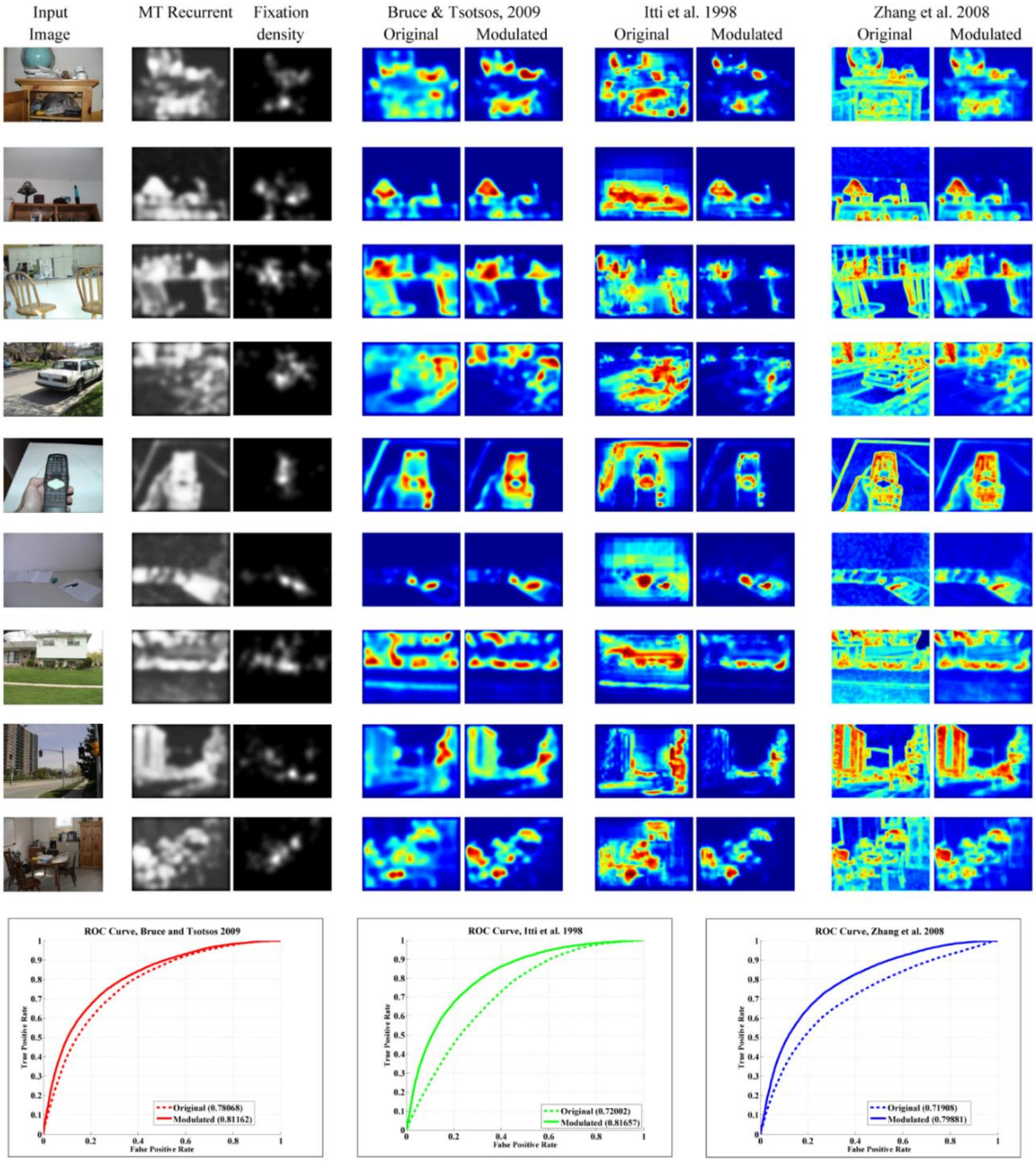


Figure 5-2 Comparison of saliency given static input images. Up part from left to right: original images, feedback strength elicited from the fast dorsal activation, visual saliency based on modulated ventral features and visual saliency based on non-modulated ventral features. Bottom part: associated ROC curves.

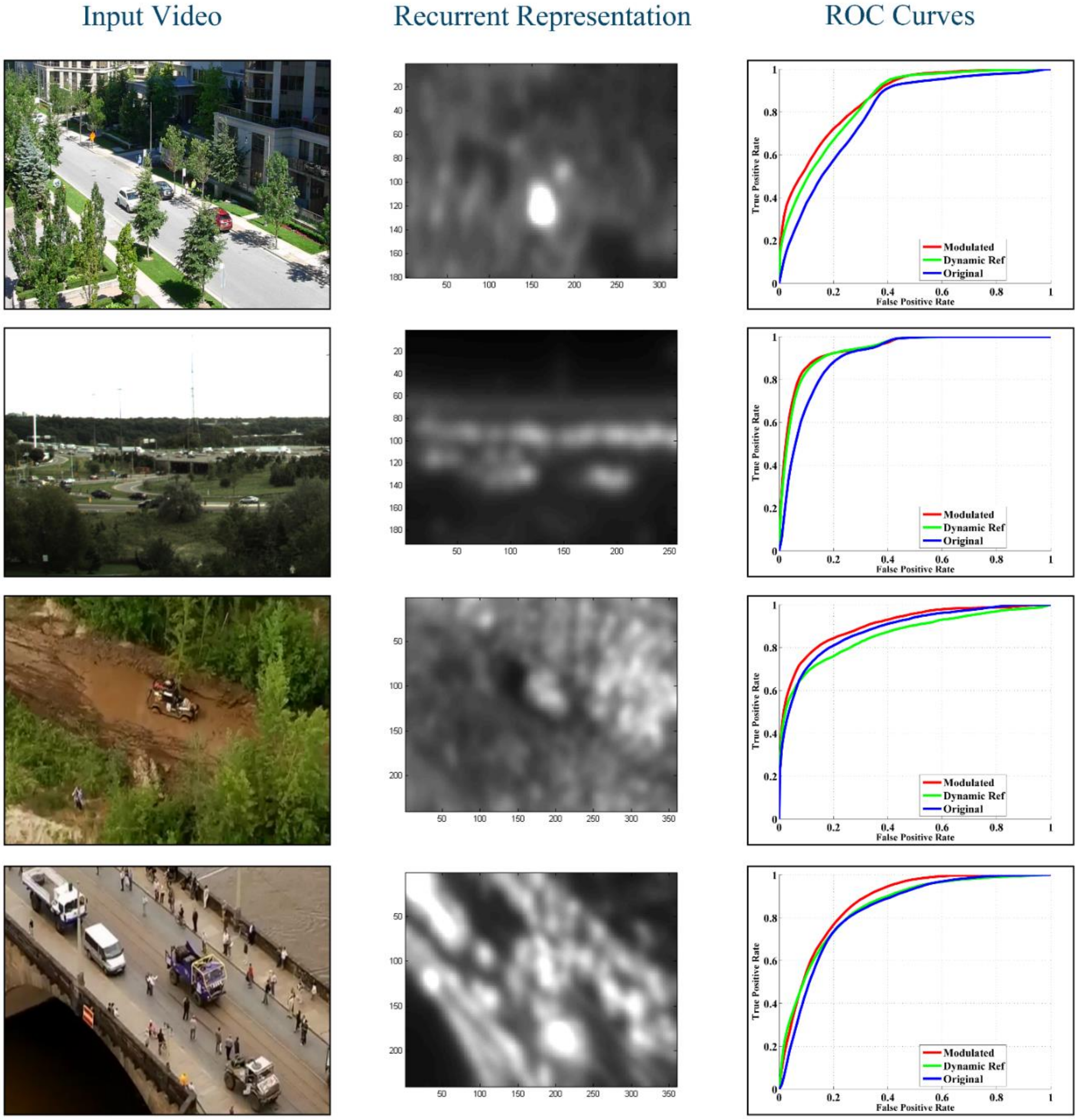


Figure 5-3 Use early spatiotemporal recurrence to improve visual saliency. Left: image from test videos. Middle: early recurrent representations that highlight regions consisting with moving objects. Right: Mean ROC curves of original (Bruce and Tsotsos 2009a) (blue), a spatiotemporal alternative (Bruce and Tsotsos 2009b) (green) and our current work (red).

5.2 Background Subtraction

Background subtraction refers to a general process of improving signal response of a target by removing interference of background pixels. It is a fundamental task in various computer vision and image processing applications.

Existing approaches attempt to solve background subtraction using methods from statistics (Cucchiara et al. 2003), density estimation (Lee 2005, Han et al. 2008), feature learning (Gao et al. 2008, Han and Davis 2012), etc.

5.2.1 Rationale

At first glance, the problem of background subtraction shares many similarities with the aforementioned visual saliency. We therefore hypothesized that visual saliency representations based on the modulated ventral representation may improve background subtraction performance.

We borrowed the idea of early recurrent (ER) processing from the primate visual system. During feature extraction, center-surround (CS) inhibition utilizes lateral connectivity to suppress the response of neighbourhood activations (Figure 3-7). Further, it is possible that the dorsal information plays a role in suppressing the ventral computation via early recurrence.

Inspired by the two types of inhibition, a computational model for unsupervised background subtraction is proposed (Figure 5-4). The model hypothesizes that the background of a dynamic scene can be eliminated in two steps. First, spatiotemporal features are computed by the modelled dorsal pathway. In this representation, activations of the foreground and background are mixed. By CS inhibition, a substantial portion of the background may be suppressed, leading to a refined spatiotemporal representation containing perceptually salient foreground only. Second, the refined spatiotemporal representation is used to inhibit the fine-scale spatial features computed by the modelled ventral pathway, such that foreground object features are accurately localized. In its most straightforward manner, ER inhibition is defined as pixel-wise multiplication.

5.2.2 Experiment

We investigated the effect of center-surround inhibition and early recurrent inhibition in subtracting background. We used real video sequences from (Gao et al. 2008), which have been widely used for background subtraction applications in the literature.

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. The computation simulates the early recurrent processing between the dorsal and the ventral pathways of the primate visual system. The process can be described as:

$$R_i(x,y) = H(E_i^V(x,y) \cdot Inh_i^{ER}(x,y)), \tag{5-1}$$

where $R_i(x,y)$ denotes the output for features i. $H(s) = \max(s,0)$ is a rectification function. $E_i^V(x,y)$ denotes energy of ventral (spatial). In many existing works, E_s^V is deemed as output representation. $Inh_i^{ER}(x,y)$ denotes early recurrent inhibition between the dorsal pathway and the ventral pathway as:

$$Inh_i^{ER}(x,y) = H(E_i^D(x,y) - \alpha \cdot Inh_i^{CS}(x,y)), \tag{5-2}$$

where $E_i^D(x,y)$ denotes energy of dorsal (spatiotemporal) features. $Inh_i^{CS}(x,y)$ denotes CS inhibition, and α is a constant that weights the CS inhibition.

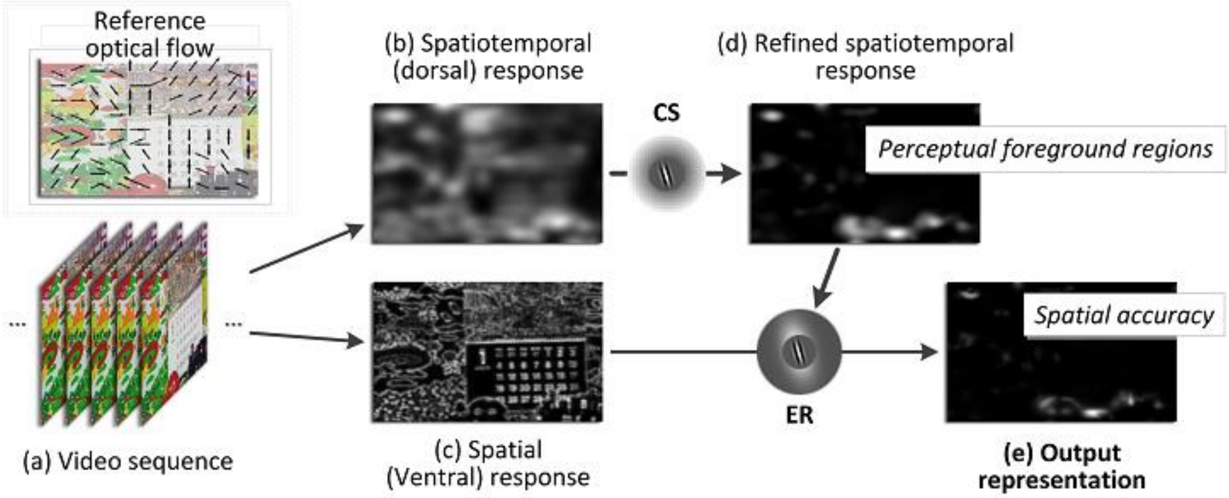

Reference optical flow
(b) Spatiotemporal (dorsal) response
(d) Refined spatiotemporal response
CS
Perceptual foreground regions
Spatial accuracy
ER
(a) Video sequence
(c) Spatial (Ventral) response
(e) Output representation

Figure 5-4 The proposed model of background subtraction using Tempete video sequence (a). Background of dynamic scene can be eliminated in two steps. 1) Spatiotemporal features are computed by the modelled dorsal pathway (b). By CS inhibition, background pixels are suppressed, leading to a refined representation containing only foreground pixels (d). 2) The refined representation inhibits the fine-scale spatial features computed by the modelled ventral pathway (c), leading to the final output representation (e).

Formalization of ventral features $E_i^V(x,y)$, dorsal features $E_i^D(x,y)$ are as proposed in Chapter 3.2. CS inhibition $Inh_i^{CS}(x,y)$, and ER inhibition $Inh_i^{ER}(x,y)$ are discussed in the rest of this section.

Center-surround inhibition weighting

The center-surround inhibition, Inh_i^{CS} is defined in an anisotropic manner to self-inhibit the dorsal representation. The process is formalized as a convolution of dorsal energy E_i^D with a weighting function, which is defined as:

$$Inh_\theta^{CS}(x,y) = E_\theta^D(x,y) * \omega_\theta^D(x,y), \tag{5-3}$$

$$\omega_\theta^D(x,y) = \frac{H(DoG_\sigma(x,y))}{\|H(DoG_\sigma(x,y))\|_1}, \tag{5-4}$$

where $DoG_\sigma(x,y)$ denotes the centre-surround strength, and $\|.\|_1$denotes the L1 norm. σ_c is the center bandwidth, and σ_s denotes the surround bandwidth. We set $\sigma_s = 4\sigma_c$ following (Kaplan et al. 1979, Einevoll and Plesser 2012).

The result of CS inhibition is a refined dorsal representation that perceptually catches the foreground, as shown in Figure 5-4 (d). Due to its low-spatial frequency response profile, this representation lacks spatial-accuracy.

Figure 5-5 illustrates the effect of CS inhibition with different α values. The first row represents the overall inhibition from the dorsal pathway $Inh_i^{ER}(x,y)$, and the second row shows the inhibited ventral features. When α increases, background pixels (wallpaper and calendar) fade out gradually, and foreground pixels (ball and train) remain mostly untouched. If α continues increasing, the foreground will also be suppressed. We noticed $Inh_i^{ER}(x,y)$ covers target regions that are perceptually salient. However, compared with the ventral representation, $Inh_i^{ER}(x,y)$ is relatively coarse.

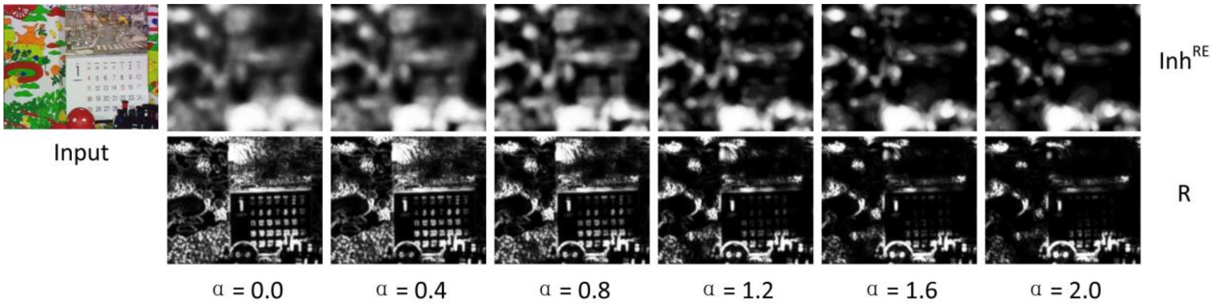


Figure 5-5 Centre-surround inhibition weighting parameter α. Motion patterns included in the input: camera motion (leftward), calendar motion (upward), ball motion (leftward) and train motion (leftward).

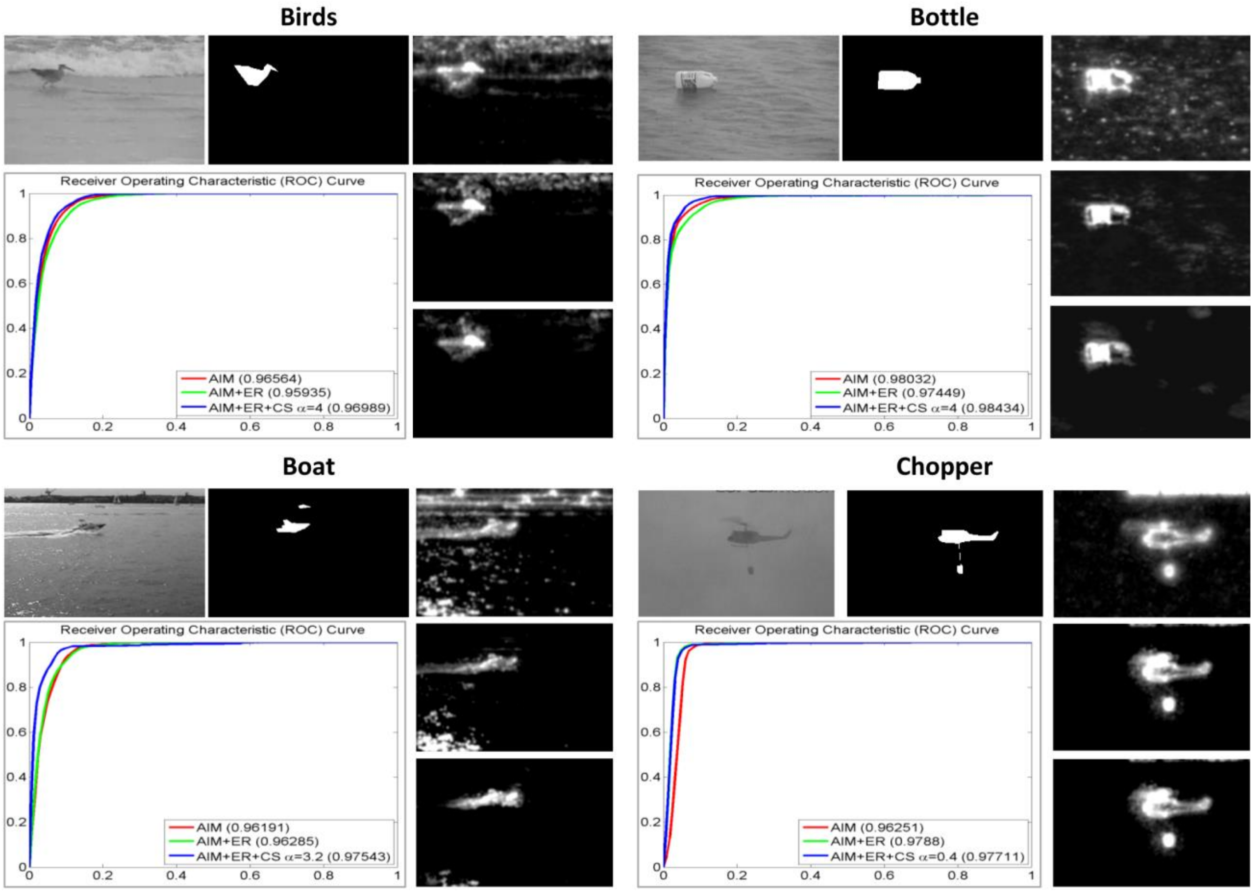


Figure 5-6 Saliency representation computed by AIM. For each sequence, figures in clock-wise order: input, ground truth, original AIM saliency, AIM+ER, and AIM+ER+CS. Also shown are the mean ROC curves over all frames for each sequence. The area under each curve is displayed in the bracket of the legend.

To quantitatively evaluate the effect of background subtraction, output is attached to a state-of-the-art saliency model, AIM (Bruce and Tsotsos 2009a) to compute visual saliency. The goal is to determine whether the feature maps refined by $Inh^{ER}(x, y)$ lead to improved saliency representations. It is thus natural to deem AIM based on original feature maps to provide baseline performance.

Real scene sequences (with ground truth) from (Gao et al. 2008) are used. This dataset contains different types of figure-background and spatiotemporal variations. It has been widely used in related background subtraction studies.

Figure 5-6 compares different saliency representations. The right column of each video illustrates saliency maps from top to bottom: based on original features (AIM), based on features modulated by ER inhibition only (AIM+ER), and based on features modulated by both CS and ER inhibitions (AIM+ER+CS). High intensity values indicate high saliency. We see that there are more similarities between the ground truth (top-middle drawing) and the salient regions computed by AIM+ER+CS.

Saliency performance is measured by ROC curves over all frames for each sequence. Given the ground truth, the ROC curve is defined as true positive rate versus false positive rate. It is clearly shown in Figure 5-6 that the curves generated by AIM+ER+CS augment the other two algorithms significantly. Area under curve (AUC) is calculated. ER+CS inhibitions raise AUC in most tests, which further confirm that early recurrence is a generally effective method in background subtraction.

In conclusion, we proposed a novel approach of unsupervised background subtraction for dynamic scenes. The model is inspired by the early recurrent processing of the primate visual system. In this model, representation computed by the dorsal pathway is perceptually consistent with foreground. Representation computed by the ventral pathway, on the other hand, is a spatially accurate description of mixed foreground and background variations. The model includes two types of inhibition, the center-surround inhibition and the early recurrent inhibition. They improve the ventral representation by inhibiting responses to background pixels. Using a saliency model, we quantitatively evaluated the subtraction performance. Results using real scenes clearly indicate that the proposed work is a generally applicable process.

5.3 Fast Scene Recognition

In view of the enhanced saliency representation, we further hypothesize that early recurrence is useful for scene recognition. To test this idea, we applied the modulated ventral representation to a back-propagation neural network as a scene classifier. Video clips introduced in (Siagian and Itti 2007) were used for our test. They include a variety of cluttered scenes and have been widely used in the same type of applications. These clips were recorded using a hand-held camera (see Figure 5-7). We wanted to examine how spatiotemporal information extracted by the dorsal representation may influence the overall scene recognition performance. Furthermore, two other recognition systems were used for comparison. They represent different feed-forward recognition strategies.

We employed a straightforward machine vision strategy for recognition. A holistic representation (ER) is implemented following (Siagian and Itti 2007). Visual features corresponding to the two visual pathways are computed. Specifically, the dorsal representation describes camera motions and coarse scene arrangements. In our implementation, there are 12 dorsal feature maps, representing 12 moving directions (with 30 degree intervals). The ventral representation catches content details. There are 12 ventral feature maps, representing 12 spatial orientations (with 30 degree intervals). Via early recurrence, the ventral feature maps are modulated. To construct a holistic representation, each modulated ventral feature representation is cut into 5-by-5 non-overlapping blocks. The average intensity value of each block is calculated. We then transformed these values into a feature vector representation by ordering them from left to right and from top to bottom. Thus, each feature vector has 25 elements. Finally, vectors of the ventral features are concatenated to form the holistic vector representation, which is then sent to the recognition network.

The other two systems use the same filters to compute visual features, but they employ different strategies to construct the holistic representation. The first one (SI) represents the feed-forward strategy introduced in (Siagian and Itti 2007), where the holistic representation is based on the non-modulated ventral features. We introduced a second feed-forward strategy as the benchmark system (BM), which builds the holistic representation by the stacking of dorsal features and ventral features. The reason for introducing the benchmark system is that it represents another way to handle motion features, similar to the dynamic reference in Section 5.1 (Bruce and Tsotsos 2009b). Thus, ER and SI both include 12 features in the holistic representation, while BM includes 24 features. The length of vector in the holistic representation for ER and SI is $5 \times 5 \times 12 = 300$, and it is $5 \times 5 \times 24 = 600$ for BM.

As shown in Figure 5-8, test sequences include 3 scenarios of a university campus, “ACB”, “AnFpark” and “FDFpark”. Each scenario includes 9 different scenes, with each scene under varied illuminating conditions. In our experiment, six clips of each scene were used to train the network, with the remaining four clips used to test performance.

We applied the same one-hidden-layer back-propagation neural network in all systems. To further simplify the computation and allow all the learning networks to have the same number of input nodes, we reduced the vector length of the holistic representation to 80 using principal component analysis available in (Hyvarinen 1999). Therefore, the input layer of the network includes 80 nodes. The output layer contains 9 nodes, with each corresponding to a scene within a scenario. The neural network contains one hidden layer of 100 nodes. For fair comparisons and to exclude the performance gain introduced outside the proposed fast recurrent modulation, all tests use the same set of network parameters.

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. Table 5-1 provides a quantitative comparison of performance achieved by the three systems to correctly recognize a scene, as per scenario. Performance is measured by recognition correctness, the ratio between the number of true positives and the number of all test samples. We see from the table that ER outperforms the other two systems for all scenarios. The empirical conclusion drawn in the previous section that early recurrent modulation is able to provide a better figure-ground segmentation to facilitate object recognition is confirmed in this quantitative study.

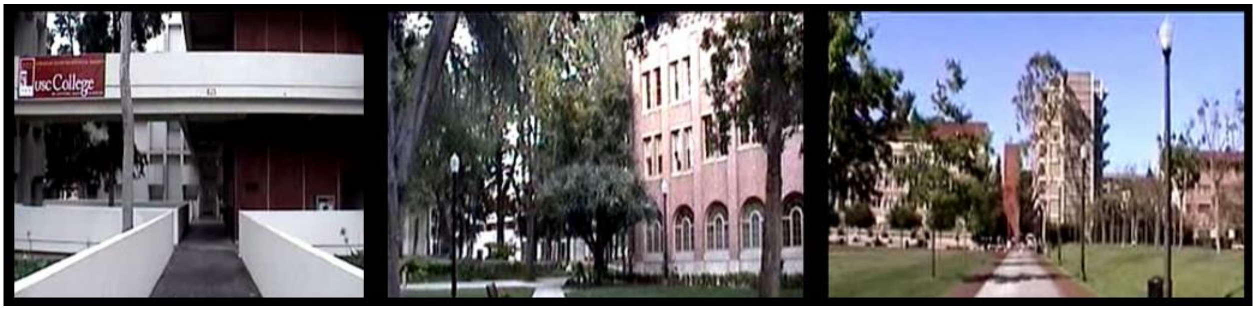


Figure 5-7 Test sequences in (Siagian and Itti 2007). From left to right: ACB, AnFpark and FDFPark.

Table 5-1 Comparison of recognition performance. The percentage indicates the correctness rate, which is computed as the number of correctly recognized scenes divided by number of total test scenes. The proposed method consistently outperforms Siagian and Itti's method and the benchmark system.

Scene	**Siagian and Itti (SI)**	**Early Recurrence (ER)**	**Benchmark System (BM)**
ACB	90.43%	93.84%	91.25%
AnFpark	90.62%	91.45%	91.22%
FDFpark	90.26%	93.16%	92.41%

5.4 Discussion

In this chapter, we applied the model of early recurrence to improve visual saliency. In its most simplified form, the model applies results computed by higher-level dorsal areas to inhibit the computation of lower-level ventral areas, such that it is the modulated ventral representation that is used to construct the saliency map.

The proposed model uses localized content influence. This is different from the existing models involving scene Gist (Oliva 2005, Oliva and Torralba 2007) in many ways. Although both models are proposed with the goal of having contextual representation affect object processing, their motivations and biological foundations are rudimentarily different. The focus of Oliva and colleagues is to use context to prime the input image with regions that are most likely to contain targets. Such a process is based on a model involving a contextual prior that learns target features and locations from experience. Contextual representation in this proposal captures the characteristics of visual inputs in a direct format that reflects the function of the neurons. In this manner, our model allows multiple contextual representations that correspond to different types of neurons in the dorsal pathway. Second, although both context models are integrated to impact image saliency, our model works on saliency improvement only to demonstrate its usability. Contextual representations may be of a variety of different natures and are computed independently from any specific saliency algorithm. In general, the existing approach follows a “whole scene” paradigm. Ours, on the other hand, is a multi-scale, image-based approach. The computation formalized in our model is consistent with biological vision that different features are computed with different speeds in the early visual system and hence can positively affect one another through fast recurrent connections.

Chapter 6. Impact of Early Recurrence on Contour Representation

Edge detection is a basic yet challenging component in various image analysis and computer vision applications. Numerous edge detection algorithms have been proposed in the past decades (Marr and Hildreth 1980, Canny 1986, Mehrotra et al. 1992, Thune et al. 1997, Grigorescu et al. 2003, Ren 2008). These models deal with the problem from a number of approaches, such as image filtering, statistical analysis, and machine learning. However, the problem of edge detection remains only partially solved. Most of these models extract reliable edges given simple input (without much scattered content or image noise), but merely effective in complicate scenarios.

Perhaps the most difficult requirement of edge detection is how to localize real edges accurately while excluding distracting pixels (fake edges) caused by image noise and fine texture. For filter-based edge detection algorithms, two factors determine the performance: filter design and discrimination method. A good filter is sensitive to edge (pixel intensity discontinuity) with accurate localization. However, the side effect is that false edges may also emerge in the computed edge representation. This situation then requires a decision-making method to discriminate real edges from false edges. In this direction, statistical analysis (Thune et al. 1997), multi-scale analysis and machine learning (Ren 2008) techniques have been investigated to train the model to be more sensitive to real edges. However, learning methods have the disadvantage of over-fitting where a trained model may favourably predict real edges given similar-to-training images, but drastically lose predictability given new or unseen inputs.

6.1 Early Recurrence Improved Edge Representation

Inspired by the biological vision, a number of edge detection models have been proposed to make use of a non-classical receptive field (RF) to facilitate edge detection. In (Grigorescu et al. 2003), the center-surround mechanism observed in visual area V1 is modelled. The mechanism manifests itself as self-inhibition that suppresses edge segments surrounding the center RF (see Figure 3-7 left drawing). The authors formalized two types of self-inhibition, isotropic inhibition (all responses outside the center RF are suppressed in an equal way, independently of their preferred orientations) and anisotropic inhibition (only responses obtained for the same preferred orientation as a central response are suppressed). Experiments using real images confirmed that self-inhibition indeed improves edge detection.

The essence of the center-surround inhibition is that via lateral connections the center response suppresses distracting pixels in the surrounding region. If both center and surround regions share similar pixel values, it is then unlikely that the center region includes an edge. On the other hand, if there is strong center-surround difference, then the center region is likely to be on an edge, or at least contains content that is different from its surroundings. In the mathematical formulation, the inhibition boosts contrast, which relatively increases the center pixel values in the edge representation. Therefore in the final edge map, stronger pixel values more likely posit real edges.

However, each center-surround filter has a preferred spatial scale. The actual performance is thus highly dependent on the spatial scale of the filter. Although self-inhibition has been extended with a multi-scale analysis (Papari et al. 2007), it is noted and will be shown in the our experiment that in complicated scenes where targets are at different spatial scales, output contours are inconsistent. Another disadvantage of this center-surround mechanism is that the surrounding region is circular shaped. If there is a real edge in the surrounding region, then its pixels values associated with the real edge are suppressed together with distractors, leading to false negatives.

Self-inhibition has been recently revisited with a refined center-surround inhibition scheme (Zeng et al. 2011). Instead of the circular-shaped region, a butterfly-shaped region is proposed. It consists of two adaptive inhibitory end-regions and two non-adaptive inhibitory side-regions. This region performs better than (Grigorescu et al. 2003) in preserving real edges. However, the filter is mathematically difficult to derive, making it unlikely to fit into a real vision system. Further, the biological underpinning to such butterfly-shaped self-inhibition region is unclear.

In this work, we proposed to use early recurrent inhibition, in addition to self-inhibition, to improve edge calculation. Motivated by the fast-brain hypothesis, we computationally formalized the early recurrent processing. The computation simulates the recurrent operation from the dorsal area MT to the ventral layers of visual area V1 (ventral V1). An important physiological difference between area MT and ventral V1 is that they respond to different spatiotemporal image features. Relatively, neurons in area MT are more sensitive to coarse-level spatial information, which contains a brief content description. Neurons in the ventral

V1 respond actively to fine spatial variations, which correspond to edges and textures. Therefore, the essence of the early recurrent operation is that edges and textures are suppressed by the brief content description. As we showed in the previous chapter, the dorsal representation has a marked correspondence to image saliency although lacks spatial accuracy. In this way, the dorsal representation informs the ventral processing with regions that are likely to contain objects of interest. By the modulation, edges and textures that do not consistent with the dorsal representation will be inhibited, leading to improved edge representation.

It is proposed that the early recurrent inhibition is a weighted multiplication operation. The right figure of Figure 3-7 visualizes the idea of the spatial region surrounding the ventral neuron. Compared with self-inhibitions (left and middle figure of Figure 3-7), the shape of the surrounding region of early recurrence depends on the dorsal neurons.

In Chapter 3, we sketched three early recurrent inhibition patterns. In edge detection, we implemented two of them, the isotropic inhibition and the anisotropic inhibition. Similar to the concept presented in (Grigorescu et al. 2003), the isotropic inhibition causes ventral responses to be inhibited by the summation of dorsal responses to all orientations in an equal manner. The anisotropic inhibition suppresses ventral V1 responses to a preferred orientation by MT responses to the same orientation.

Our proposal is consistent with the scale-space theory (Witkin 1983) and suggests that early recurrence could play a role in enforcing consistency of image structure across scales. In our model, we implemented a strategy that computes the dorsal representation using the interval-tree technique proposed in (Witkin 1983). In the scale-space, the algorithm calculates a coarse-scale representation by searching for the local maximum stability covering. The covering highlights a marked correspondence between the stability of an object's contour and its perceptual salience. Fine-scale representation, on the other hand, is a highly-spatially accurate edge map. By inhibiting the fine-scale edge map with the coarse-scale representation, a substantial number of false edges caused by image noise and texture variations may be removed; leaving real edges in line with coarse-scale representation remain.

To investigate the impact of early recurrence, we used refined edge representations generated by the two inhibition methods as inputs to a contour operator used in (Grigorescu et al. 2003). Using real images, we quantitatively compared contours calculated by our work with the contour detector proposed in (Grigorescu et al. 2003). Note that method in (Grigorescu et al. 2003) is also a biologically motivated model. In order to compare our model with non-biologically motivated works, in another experiment, we applied early recurrence to boost existing contour extraction models. Results from both experiments clearly demonstrate that early recurrence has a positive and consistent impact on contour detection. Further, we showed that coarse edge representation calculated via scale-space analysis achieves the best performance.

6.2 Implementation

We modelled the two visual pathways to compute different edges. Since the two pathways start from area V1, we used the term ventral V1 (V1v) and dorsal V1 (V1d) to refer to layers in area V1 that are abstractly associated with the two pathways. Specifically, V1v computes fine-scale edges and projects results to higher-level ventral areas to compute object contours. V1d is sensitive to coarse-scale edges and sends output to area MT for further integration.

Feed-forward formalization of each component has been derived based on Chapter 3.2.

The result of area MT is sent back to modulate V1v computation (Figure 6-1 left drawing). The modulated V1v representation is further inhibited by the center-surround self-inhibition (Grigorescu et al. 2003). The whole process is defined as:

$$R^{V1v}(x,y) = H(E^{V1v}(x,y) \cdot Inh_{er}^{MT}(x,y) - \alpha Inh_{cs}^{V1v}(x,y)), \qquad (6\text{-}1)$$

where R^{V1v} denotes modulated V1v representation. $H(s) = \max(s, 0)$ is a half-wave rectification function. E^{V1v} denotes V1v responses to feed-forward image stimuli. Inh_{er}^{MT} is the early recurrent inhibition generated from area MT. Inh_{cs}^{V1v} denotes the center-surround self-inhibition, and α is a weighting factor. Figure 6-1 shows the overall structure of the proposed model.

We proposed Inh_{er}^{MT} inhibits V1v via multiplication. Further, Inh_{cs}^{V1v} is defined as isotropic non-classical RF inhibition (Grigorescu et al. 2003). Note that although self-inhibition is generated within area V1v, the temporal aspects of the asynchronous signal projection properties between the dorsal and the ventral pathways make it possible that the recurrent inhibition from area MT impacts ventral V1v prior to its self-inhibition (see the temporal requirement study in Chapter 3.1).

To simplify the work, the 3-dimensional scale-space is separated into two 2-dimensional scenarios, $x - q$ and $y - q$. They are analyzed separately. To reduce computation, only two gradients of Gaussian (0° and 90°) are used. They are defined as:

$$\nabla_x g^{MT}(x, y, \sigma_{gx}) = \left(-\frac{x}{\pi\sigma_{gx}^2}\right) e^{-(x^2+r^2y^2)/(2\sigma_{gx}^2)}, \qquad (6\text{-}2)$$

$$\nabla_y g^{MT}(x, y, \sigma_{gy}) = \left(-\frac{x}{\pi\sigma_{gy}^2}\right) e^{-(x^2+r^2y^2)/(2\sigma_{gy}^2)}, \qquad (6\text{-}3)$$

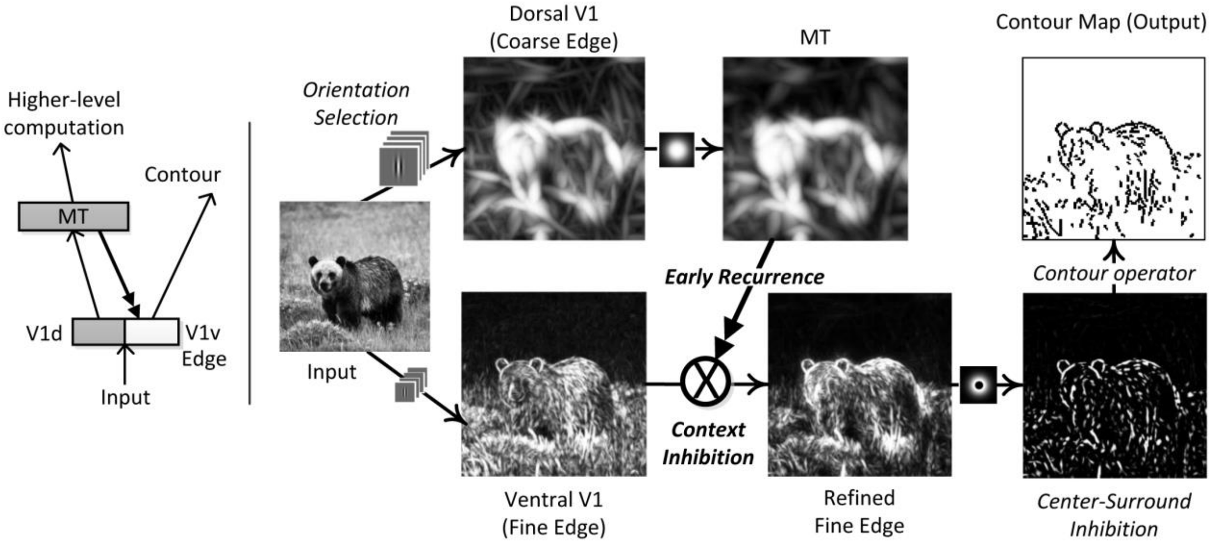


Figure 6-1 General structure of the proposed model. Left: simplified biological hierarchy and connections. The double arrow line from MT to V1v denotes early recurrence. Right: an example of using the proposed computational model to compute object contours.

To simplify the work, the 3-dimensional scale-space is separated into two 2-dimensional scenarios, $x - q$ and $y - q$. They are analyzed separately. To reduce computation, only two gradients of Gaussian (0° and 90°) are used. They are defined as:

By increasing σ_{gx} and σ_{gy}, scale-space representations $x - q$ and $y - q$ are constructed respectively. As shown in Figure 6-2, given a $x - q$ representation, an interval-tree is built. It is observed that scales of real object contours have a marked correspondence with stability (vertical axis) in the scale-space. Therefore, the algorithm searches for a covering of the space (a set of gray blocks), which includes the most stable blocks across all intervals (horizontal axis). Here, stability is indicated by the scale spans. It is such that the larger the scale in each block, the more stable that scale. The algorithm then selects the largest scale $\hat{\sigma}_{gx}$ that crosses the most stable blocks in the covering as the output scale at each spatial location, and the output of MT of 0° is formalized as:

$$r^{MT}(x, y, 0) = \nabla_x g^{MT}(x, y, \hat{\sigma}_{gx}), \tag{6-4}$$

The same operation also applies to the $y - q$ space.

Early Recurrence

We proposed that the early recurrent inhibition between area MT and ventral V1 can be represented as a weighted multiplicative process. The inhibition representation is defined as:

$$Inh_{er}^{MT}(x, y, 0) = \frac{\sum_{\delta \in \Delta} \omega(\delta, \theta) r^{MT}(x, y, \theta)}{\left\| \sum_{\delta \in \Delta} \omega(\delta, \theta) r^{MT}(x, y, \theta) \right\|_1}, \tag{6-5}$$

where $\omega(\delta, \theta)$ is the weighting factor, denoting the strength of connection between MT neuron of orientation δ and ventral V1 neuron of orientation θ. Σ denotes the summation of MT neurons for all orientations $\delta \in \Delta$. $\|.\|_1$ is the L1 norm. By setting $\omega(\delta, \theta)$, two special types of early recurrent inhibition scheme are derived and investigated separately.

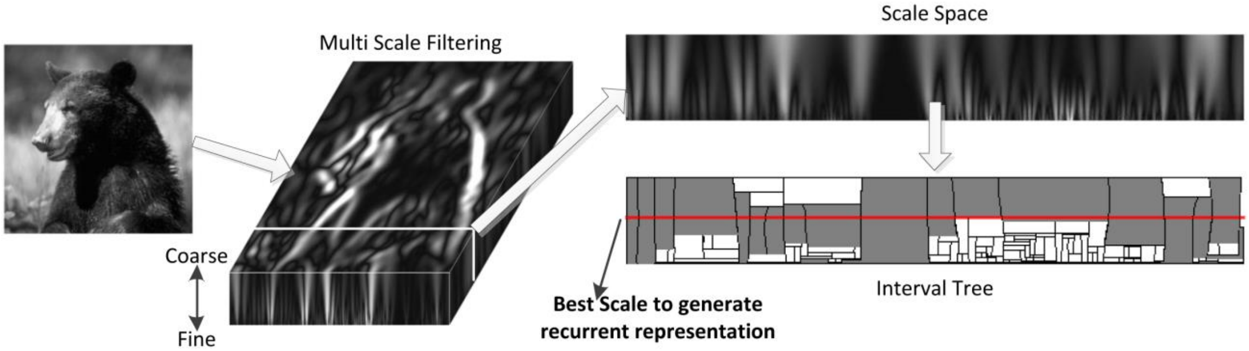


Figure 6-2 Scale-Space Analysis. Given an input image, the scale-space representation is computed coarse-to-fine. To simplify the work, the three-dimensional scale-space x-y-q is separated into x-q and y-q respectively. An interval tree is built following (Witkin 1983), based on which the best coarse-scale representation is determined via searching for a covering of the space (gray blocks).

R1. Isotropic inhibition causes fine-scale edges of an orientation to be inhibited by MT with all orientations in an equal manner. To do this, we fixed $\omega(\delta, \theta) = 1$ for all orientations. The isotropic representation is a summation of MT. It highlights regions corresponding to low spatial frequency variations to all orientations and is insensitive to variations caused by high spatial frequency stimuli (i.e., noise and textures).

R2. Anisotropic inhibition suppresses fine-scale edges to by MT responses of the same orientation:

$$\omega(\delta, \theta) = \begin{cases} 1, & if\ \delta = \theta \\ 0, & otherwise \end{cases}, \tag{6-6}$$

Each MT representation contains information of low-spatial frequency variations to only one orientation. It then modulates V1 responses to the same orientation.

Center-Surround Self-inhibition

Center-surround isotropic self-inhibition, $Inh_{cs}^{V1v}(x, y)$ is formalized following [1] as a convolution of the maximum energy map $\widehat{\mathrm{E}}^{\mathrm{V1}}(\mathrm{x}, \mathrm{y})$ with a weighting function as:

$$Inh_{cs}^{V1v}(x, y) = \hat{E}^{V1}(x, y) * \omega^{V1v}(x, y), \tag{6-7}$$

where the maximum energy map is calculated by finding the maximized filter responses among N orientations.:

$$\hat{E}^{V1}(x, y) = \max\{E^{V1}(x, y, \theta_i) \mid i = 1..N\}, \tag{6-8}$$

And the weighting function is defined as a Difference of Gaussian function:

$$\omega^{V1v}(x, y) = \frac{H(DoG(x,y))}{\|H(DoG(x,y))\|_1}, \tag{6-9}$$

6.3 Experiment

The proposed computation has been implemented using Matlab. The simulation is conducted on a Windows 7 PC. We implemented the proposed model in three forms: (fI) isotropic inhibition based on single-scale recurrence, (fA) anisotropic inhibition based on single-scale recurrence, and (ssI) isotropic inhibition based on scale-space recurrence.

A dataset of 40 images with ground truth contours are used for evaluation. This dataset has been widely used in the literature of contour detection. Although there are other datasets for edge detection, images presented in the selected one cover a broad range of spatial frequency variations, different types of textures, and different types of artifacts.

In the first experiment, we compared our model with the self-inhibition edge detector (Grigorescu et al. 2003). For this purpose, their method of isotropic center-surround self-inhibition (S) has been implemented. Detected contours are compared against ground-truth contours. For a given ground truth map (G) and a detected contour map (D), true positive (TP) pixels, false positive (FP) pixels, and false negative (FN) pixels are those marked. The performance measurement introduced in (Grigorescu et al. 2003) is employed. The three criteria are specified as follows:

1) False Positive Rate:

$$eFP = \frac{card(FP)}{card(D)}, \tag{6-10}$$

2) False Negative Rate:

$$eFN = \frac{card(FN)}{card(G)}, \tag{6-11}$$

3) Performance score:

$$P = \frac{card(D)}{card(D)+card(FP)+card(FN)}, \tag{6-12}$$

where $card(s)$ denotes number of elements in set s. A lower score of eFP represents a better suppression of false edges, while a lower score of eFN denotes a better preservation of true edges. A higher overall score P corresponds to a better overall performance. Each edge detector is tested with different combinations of parameters to investigate performance.

Figure 6-3 compares the best contours of four selected images. These images contain objects of different types. For example, *gnu* contains multiple objects of different scales. *rino* and *elephant_2* include single object with rigid contours, and the background consists of high-frequency edges. The target in *bear* has hairy contours. From the comparisons of contour continuity, edge detail and background texture inhibition, it is clear that the contour maps computed by the proposed detectors achieve better performance than method from (Grigorescu et al. 2003). Contours generated by ssI are in the best agreement with ground truth.

Table 6-1 lists parameters used that lead to the best performance for images shown in Figure 6-3. It also summaries the false positive rate and false negative rate associated with the best performance scores.

Our proposed methods surpass the competitor in most performance measurements. The enhancements over self-inhibition detector can be as much as 167%. The proposed isotropic inhibition achieves the lowest false positive rate, indicating a better suppression of false edges. The proposed anisotropic inhibition has the lowest false negative rate, which further confirms it with a superior ability to retain real object contours. The best ratio between coarse level and fine level varies for different test images. Coarse-scale representation in ssI automatically selected by scale-space analysis achieves the most robust performance in all cases. In some cases, the best performance is achieved without self-inhibition.

Using box-and-whisker plots, Figure 6-4 illustrates the P scores. Note that the best performance (top bar) and median performance (red line in box) of the proposed three detectors are consistently higher than S. Due to the single coarse scale, early recurrent inhibition of fA and fI are not as stable as ssI that is based on scale-space analysis. ssI detector does not perform the best in all tests. We suspect this is caused by using gradient of Gaussian, where only horizontal and vertical orientations are considered, as opposed to other filters, e.g., Gabor filter.

To further investigate whether the proposed early recurrent model is a generally applicable process, we incorporated the scale-space analysis version of recurrent representation into two popular edge detectors. In particular, Canny edge detector (Canny 1986) is chosen because it is currently the most widely used edge detector, and the multi-scale Brightness/Texture Gradients (BTG) detector proposed in (Martin et al. 2004) is selected because it represents the state-of-the-art. Although Malik and his colleagues have extended the work (Ren 2008, Arbelaez et al. 2011), the goals of these were from learning perspectives to detect contours, and thus they do not relate to the current purpose. We use the original model implementations to provide baseline performance. The proposed recurrent operation fits itself into these models easily by modulating the original edge responses. One can then evaluate the effects of early recurrence by comparing results using modulated edge responses and using the original edge responses.

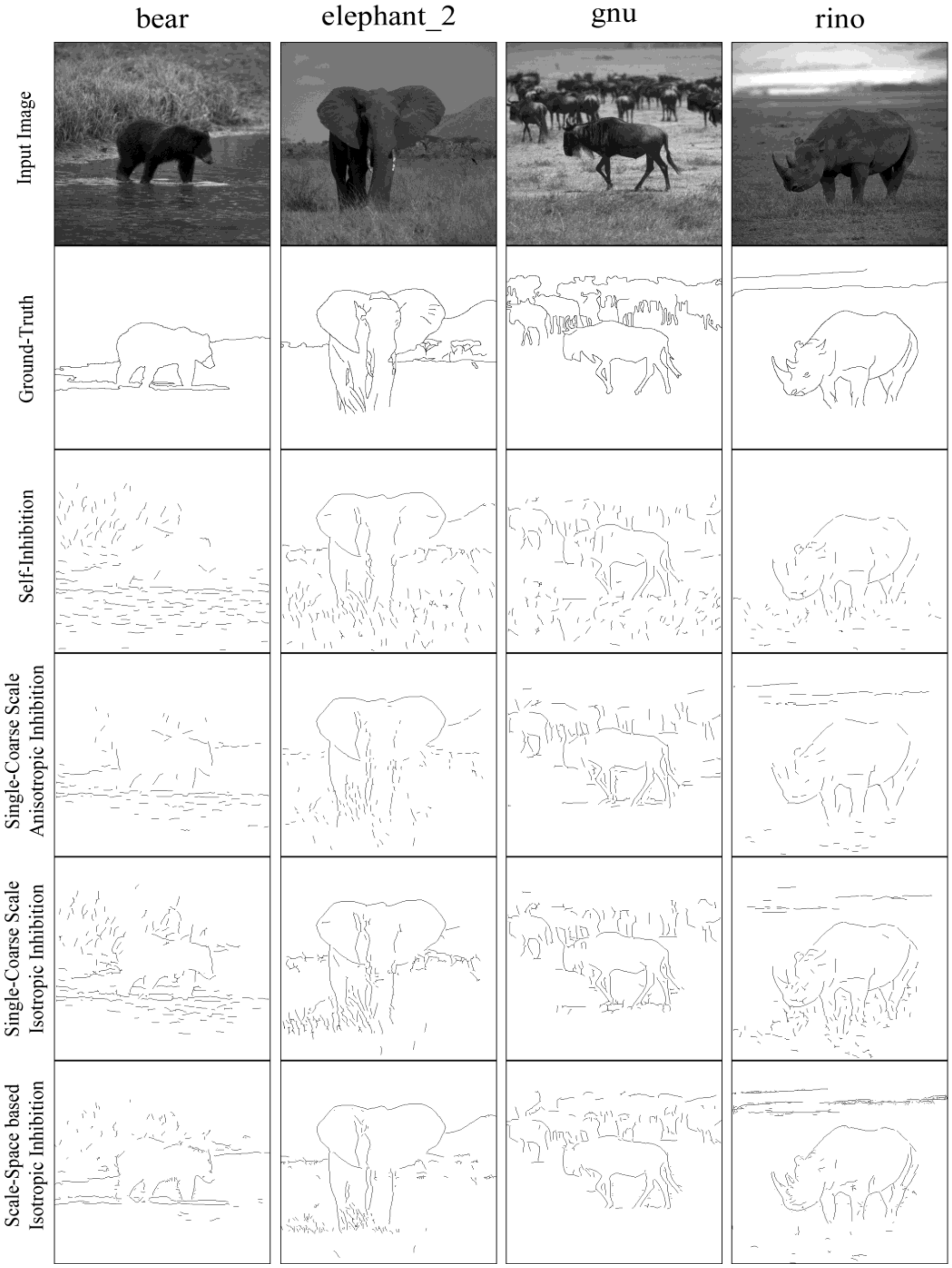
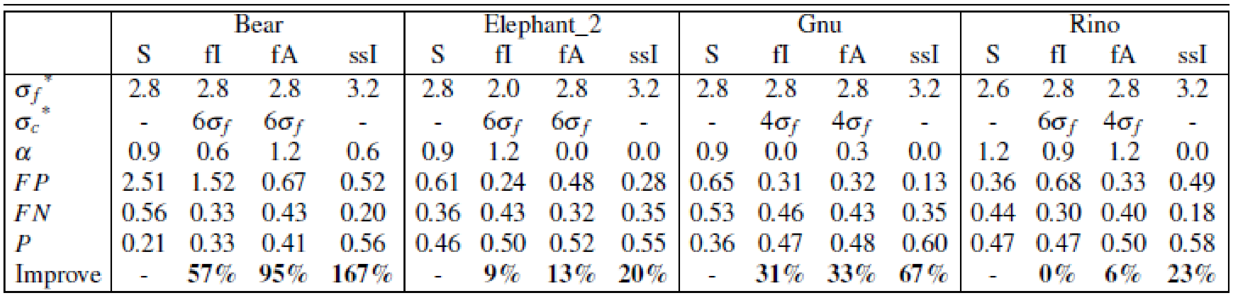

bear
elephant_2
gnu
rino
Input Image
Ground-Truth
Self-Inhibition
Single-Coarse Scale Anisotropic Inhibition
Single-Coarse Scale Isotropic Inhibition
Scale-Space based Isotropic Inhibition

Figure 6-3 Contours comparison from top to bottom, input image, ground truth, computed contours by (Grigorescu et al. 2003) and scale-space analysis ssI.

Table 6-1 Experiment parameters and performance for the images presented in Figure 6-3. Last row highlights P score improvements over (Grigorescu et al. 2003).



	Bear				Elephant_2				Gnu				Rino			
	S	fI	fA	ssI	S	fI	fA	ssI	S	fI	fA	ssI	S	fI	fA	ssI
σ_f^*	2.8	2.8	2.8	3.2	2.8	2.0	2.8	3.2	2.8	2.8	2.8	3.2	2.6	2.8	2.8	3.2
σ_c^*	-	$6\sigma_f$	$6\sigma_f$	-	-	$6\sigma_f$	$6\sigma_f$	-	-	$4\sigma_f$	$4\sigma_f$	-	-	$6\sigma_f$	$4\sigma_f$	-
α	0.9	0.6	1.2	0.6	0.9	1.2	0.0	0.0	0.9	0.0	0.3	0.0	1.2	0.9	1.2	0.0
FP	2.51	1.52	0.67	0.52	0.61	0.24	0.48	0.28	0.65	0.31	0.32	0.13	0.36	0.68	0.33	0.49
FN	0.56	0.33	0.43	0.20	0.36	0.43	0.32	0.35	0.53	0.46	0.43	0.35	0.44	0.30	0.40	0.18
P	0.21	0.33	0.41	0.56	0.46	0.50	0.52	0.55	0.36	0.47	0.48	0.60	0.47	0.47	0.50	0.58
Improve	-	**57%**	**95%**	**167%**	-	**9%**	**13%**	**20%**	-	**31%**	**33%**	**67%**	-	**0%**	**6%**	**23%**

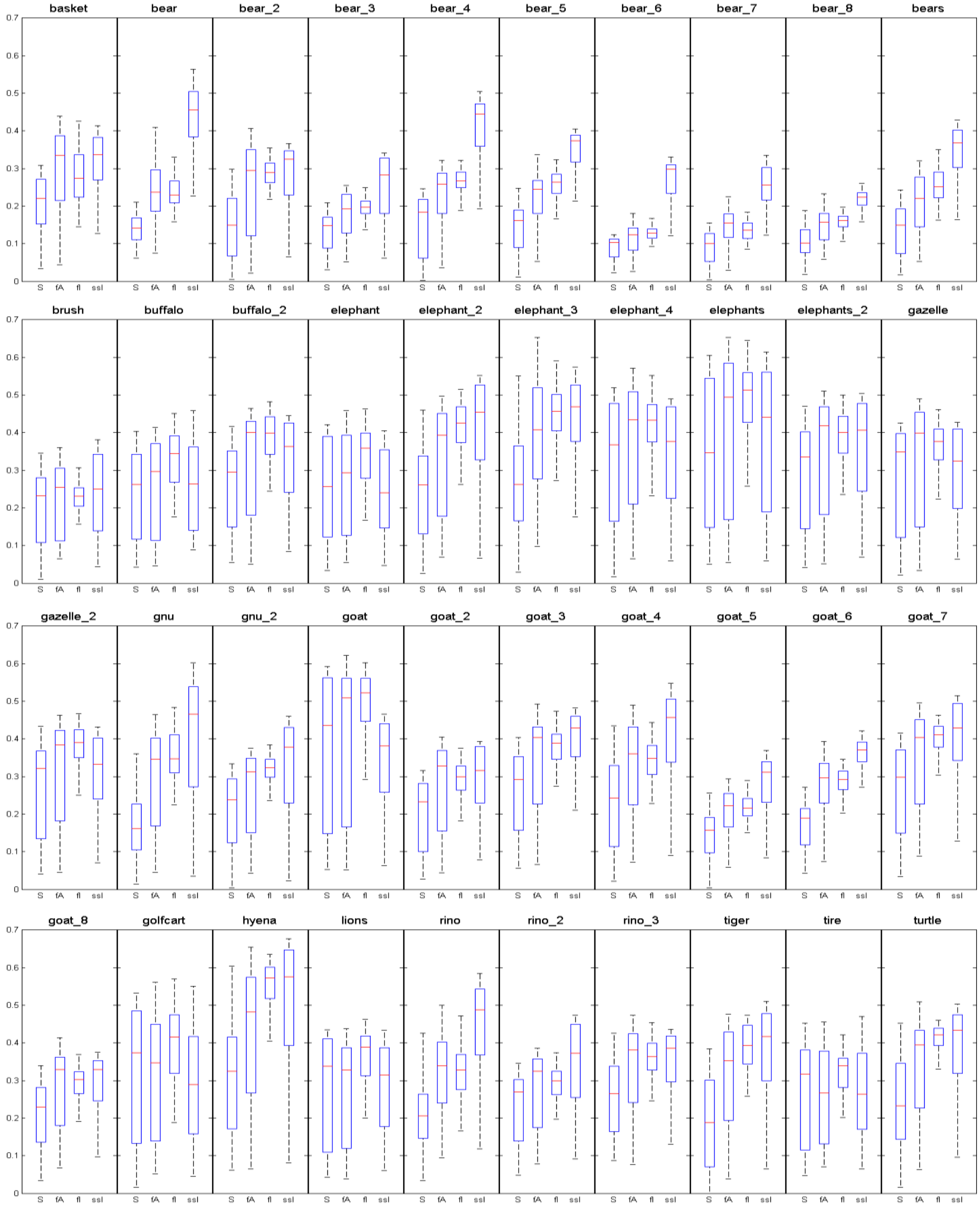


Figure 6-4 Box-and-whistler plot of all images. The proposed detectors outperform S in most cases, with ssI provideing the most reliable contours.

We tested these methods using the same dataset (Figure 6-5). Modulated by early recurrence, edges detected by both methods are more focused on the real contours, leading to much more cleaned edge representations. ROC curves and precision-recall curves are generated based on ground truth.From the comparison, we observed that curves produced from the modulated edges augment the original detectors in most cases. These results, together with the results of the first experiment, consistently indicate that early recurrence is important to lower-level feature extraction. The proposed computational model consistently improves edge detections in real scenes.

6.4 Discussion

In this chapter, we applied the proposed computational model of early recurrence to improve edge detection, which is an important application in computer vision. The essence of the computation is to use dorsal (coarse) edges to inhibit ventral (fine) edges by multiplicative inhibition.The proposed recurrent inhibition is fundamentally different from the center-surround self-inhibition. The center-surround self-inhibition is a mechanism within area V1, which works in an additive manner. Alternatively, the proposed early recurrent inhibition uses information generated from the dorsal area MT to inhibit the ventral edge representations.Experiments using real images indicate that contour maps generated by the proposed method consistently surpass self-inhibition, with respect to suppressing noise and distracting image textures, and preserving smooth contours. The isotropic and the anisotropic inhibitions are two weighting schemes studied. They facilitate edge detection differently. The anisotropic inhibition provides better output for most test images. Additional experiments are required to further explore these schemes as well as other weighting methods.

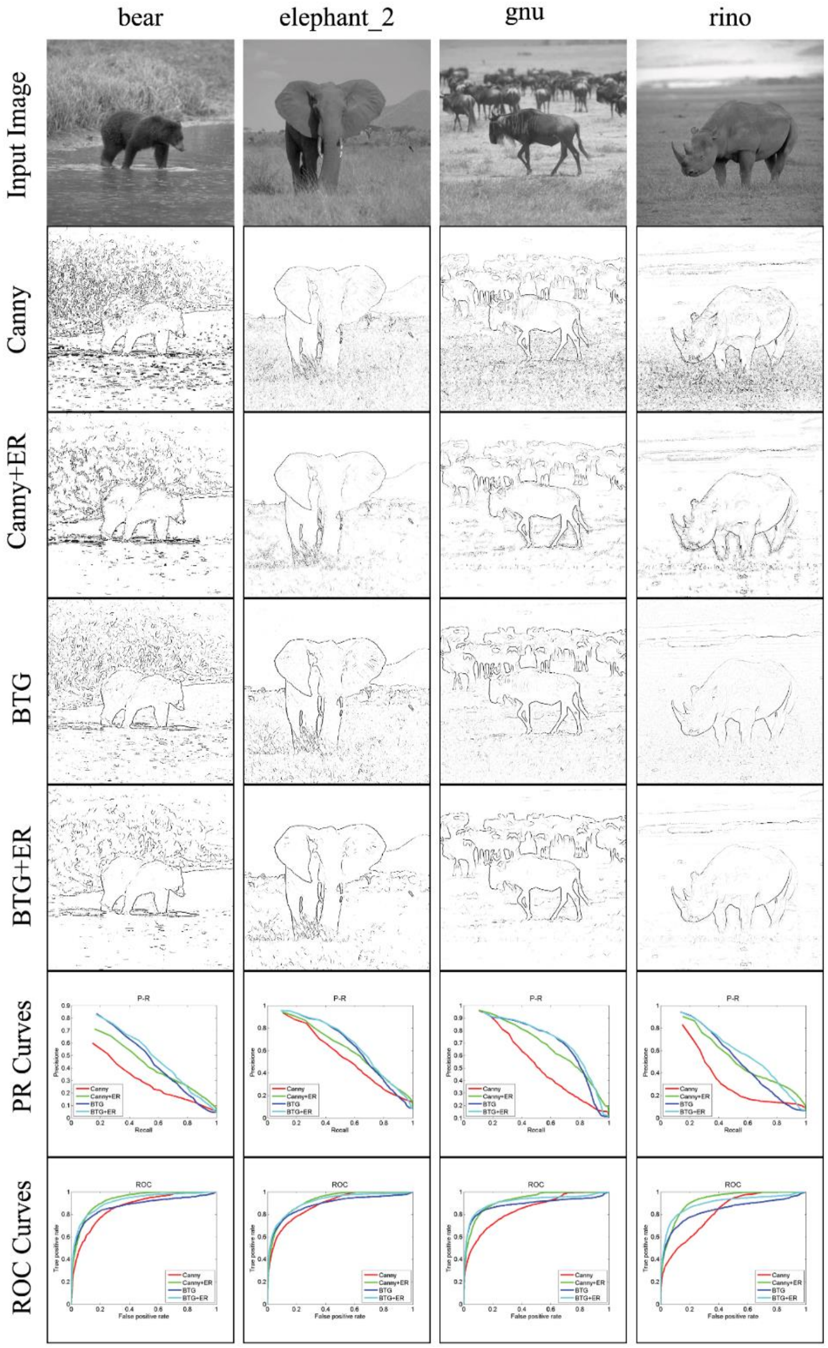


Figure 6-5 Improved contour detectors. Contours detected by early recurrent modulation (Canny+ER and BTG+ER) are significantly improved, indicating the proposed work is a general approach to boost edge detection.

Chapter 7. General Conclusions and Discussion

7.1 General Conclusions

In this dissertation, we presented a computational model to investigate lower-level recurrent mechanisms in the primate visual system. The main motivation is from recent biological studies, which indicate visual processing within the early visual hierarchy does not confine to the feed-forward hierarchy proposed by Hubel and Wiesel during the 1960s (Hubel 1963). The recent literature suggests that lower-level recurrence exists and impacts visual processing with very short temporal delays (Bullier 2001).

To answer the question of what is required for early recurrence to be at play, we conducted a literature survey to find clues. We identified two requirements. First, it requires the existence of feedback connectivity from a dorsal area to a ventral area. Based on the literature, such recurrent connections have been found among multiple visual areas (Felleman and Van Essen 1991, Lienhart and Maydt 2002).

The second requirement is that the timing of information processing must allow the feedback signals to reach the ventral neuron before the ventral neuron receives feed-forward signals from its lower-level visual area. For a long time, only cell-recording experiments had the temporal accuracy to measure latencies of visual processing from one area to another. In recent years, developments of imaging and magnetic techniques provide us non-invasive methods to learn about vision. They have provided us great insights into the asynchronized information pathways.

In Chapter 3, we followed (Schmolesky et al. 1998, Bullier 2001) to study the temporal relationships along the visual pathways, and concluded recurrent connections satisfying the above two requirements. In this work, we focused on two recurrent paths: early recurrence between the dorsal area MT and the ventral layers of area V1/V2, and early recurrence between the dorsal area MST and the ventral layers of area V2.

To answer the question of how early recurrence facilitates the ventral processing, we proposed that the essence of early recurrence is a surround suppression mechanism. It utilizes results from the dorsal pathway to improve the ventral visual representation. This is consistent with the theory of Selective Tuning (Tsotsos et al. 1995, Tsotsos 2011). Although Selective Tuning was proposed for visual attention, the model's motivation to include surround suppression and its core concept in formalizing computation are in line with early recurrence. From this aspect, the proposed early recurrent model may be seen as an additional element to the big picture of Selective Tuning.

Specifically, we proposed that the early recurrent modulation operates in a non-linear, weighted and multiplicative fashion. The non-linearity comes from both neural saturation and the fact that recurrence cannot cause neural response to change without feed-forward activations. Note that the current work does not intend to characterize the full scope of neural behaviours. Instead, we explored the potential ways in what the modulation might operate, and investigated their impacts on computer vision algorithms.

We proposed that the weighting depends on the actual recurrent connectivity. In our representation, each ventral neuron is connected with multiple dorsal neurons. These dorsal neurons may respond to different types of visual characteristics. If one type of visual characteristics dominates the recurrent strength where other types are negligible, then we termed this kind of recurrence as anisotropic recurrence. Alternatively, if all types of visual characteristics equally contribute to the recurrent strength, then we termed this kind of recurrence as isotropic recurrence. A more generalized case is that different types of visual characteristics contribute proportionally. However, it is not clear to us from the literature what the optimal way to describe this generalized case is.

Based on the simplified two-pathway visual hierarchy, we formalized the necessary computational components to implement early recurrence. We followed the filter-based approach to model a set of related subcortical and cortical visual areas. Filter parameters are determined in ways to be consistent with the receptive field properties surveyed in the literature. As such, we put forth a complete computational hierarchy. Using a synthetic image, we demonstrated that the dorsal pathway has an output representation of object motion in coarse spatial-scales, and the ventral pathway has an output representation of static edge, corner and end-stopped features in fine spatial-scales.

To investigate whether the proposed recurrent operation is biologically consistent, we simulated two well-known experiments that support early recurrence: a figure-background segregation experiment (Hupé et al. 1998) and a Kanisza illusory rectangle experiment (Seghier et al. 2000b). Our results correlate to these studies, which give us confidence that the proposed computation is capable of realizing early recurrence.

An important goal of this study is to show that early recurrence is, in general, beneficial to computer vision. To do this, we used visual saliency and contour detection as two examples to highlight the impacts. By comparing with the state-of-the-art proposals, we concluded that early recurrence can effectively improve early visual representations, which is the foundation of classic computer vision approaches. In addition, we showed that combining the proposed computation with existing algorithms is straightforward, which requires very limited implementation efforts.

7.2 Connections to Other Context Model

We stated that the early recurrent representation from the dorsal pathway takes form as localized context. It provides the ventral processing with spatiotemporal (motion) context information that is not computable in the ventral pathway.

Of course, we are not the first to attribute the dorsal representation as context. In one related model (Bar 2004, Oliva and Torralba 2007), the fast magnocellular processing is the key to generate a global view to describe the scene with objects likely within it. This global context then forms predictions to facilitate object recognition.

We have compared this global context model with the proposed early recurrence in Chapter 3. Although the two models have fundamental differences, the asynchronized visual processing of the primate visual system is a common motivation to both. This perhaps initiates a clear direction to improve existing computer vision systems.

7.3 Connections and Implications to Computer Vision

Classic computer vision systems usually start from extracting visual features from visual input, which shares many commonalities with the early stage processing of the primate visual system. This biological consistency inspires us to apply the early recurrent mechanism to improve computer vision systems. To those systems that do not begin with feature extraction, it would be difficult to directly apply early recurrence.

The current work suggests an approach to improve feature representations. To this aspect, early recurrence is closely related to the scale-space analysis in computer vision. The proposed visual hierarchy is analogous to the scale-space hierarchy to some degree. However, in our model, the layered representations do not require visual features to be the same type. Our goal is a robust scale-invariant visual representation. The biological evolution over millions of years has shaped the primate visual system with many specified mechanisms. An important goal is to attend to the most important (relevant) spatiotemporal scale from scattered world with minimal effort. We believe there is a good chance that early recurrence is one of these mechanisms that play an active role from very early and lower-level stages of the visual processing.

Early recurrence from the dorsal pathway is the key ingredient to achieve scale adaptation. Specifically, as we showed in Chapter 5 and Chapter 6, the dorsal representation has a marked correspondence with perceptually salient regions. Via recurrent inhibition, ventral neural responses not correlating to the dorsal representation are suppressed. In other words, only signals of scales that are relevant to the salient regions remain in the modulated ventral representation. Although the dorsal representation lacks spatial accuracy, it is sufficient to facilitate the ventral processing.

Compared to representations in the scale-space analysis, the proposed visual hierarchy is formalized to support different of visual features. Since the dorsal pathway has a spatiotemporal response profile, its representation naturally contains motion information. This property hints at a strategy of applying motion cues to facilitate edge response different from traditional ways. Take the scene recognition in Chapter 5 for example, where we showed that motion feature plays a positive and consistent role that influences scene recognition performance. In the comparison, the benchmark system represents a traditional way to use motion cues, as a regular feature. Firstly, we showed that this traditional way indeed has its validity to improve recognition performance. Secondly, we compared it with our strategy that uses motion as the early recurrent representation. Experiments suggested that our strategy yielded even better performance. The same idea has also been expressed in the spatiotemporal saliency experiment in Chapter 5.

In Chapter 6 we focused on the spatial aspect of the dorsal pathway. To many computer vision systems, edge detection is a fundamental task. We showed that early recurrence is capable of strengthening existing edge detection methods in complicated and scattered scenes.

From the two examples, we hypothesize that early recurrence may benefit other computer vision applications in similar ways. We showed that early recurrence is conceptually straightforward and easy to implement. More importantly, the computational cost of early recurrence is lightweight, compared with massive learning algorithms that have been widely discussed in the literature.

7.4 Future Research

The current work is motivated by the fast-brain hypothesis. Its original manuscript (Bullier 2001) describes the lower-level ventral areas (i.e., areas V1 and V2) as “active blackboards” that receive retro-injected information from the dorsal pathway to support visual processing in the higher-level ventral areas. In this work, our first intention is to implement such recurrent mechanisms with a set of principled computational components.

We further tested that our work is consistent with Bullier's initial thoughts, by simulating the bar-moving experiment (Hupé et al. 1998) and the Kanisza illusion experiment (Seghier et al. 2000a). We followed such a path to build a biologically inspired computational model. In this dissertation, we demonstrated that early recurrence is capable of improving early visual representation.

No model is without limitations. The current work has only focused on the most straightforward early recurrent connections, their properties and related inhibitions. It may be improved in many ways, and its potential usability must be explored.

One possible research direction is the inhibition mechanism. Although we have shown the impact of early recurrence, we believe the inhibition strength to refine the ventral representation may come from mechanisms other than the modelled early recurrence. For example, lateral connection is a candidate to cause similar surround suppression results. In order to distinguish a lateral mechanism from a recurrent mechanism, we may start from the timing of connection and the spatiotemporal properties of the inhibition representation. Following this hint, a comparative study would be considered beneficial to provide quantitative and qualitative differentiations between the two mechanisms.

The current work hypothesizes that early recurrence operates in a multiplicative manner. We have conducted an experiment to compare the multiplicative inhibition and an additive inhibition operation. Results seemed to support that the multiplicative operation is more consistent with the existing biological observations. This, however, may be challenged by other forms of operations, or challenged by other experiment configurations. For example, we have shown in Chapter 4 that the additive operation may also suffice to represent early recurrence in some cases.

Another interesting topic is the weighting strategy. In Chapter 3, three weighting strategies have been discussed: isotropic modulation, anisotropic modulation, and a general form. Isotropic modulation and anisotropic modulation are two extreme cases: isotropic modulation is that the ventral neuron is modulated by dorsal neurons in an equal manner, and anisotropic modulation is that each ventral neuron is modulated by dorsal neurons of the same spatiotemporal preference. Simulations and applications have been discussed in this work regarding these two cases. Future work may also explore the formulation of the general form.

Last but not least, there are many promising research topics that are closely related to the current research, including exploring more possibilities to use early recurrence in computer vision and machine vision, and bringing more insights from computational neuroscience, psychophysics and neurobiology to strengthen our understanding of early recurrence.

Reference

Adelson EH, Bergen JR. 1985. Spatiotemporal Energy Models for the Perception of Motion. J Opt Soc Am A 2: 284-299

Ahmad S. 1991. Visit: An Efficient Computational Model of Human Visual Attention. University of Illinois at Urbana-Champaign

Alonso JM, Cudeiro J, Perez R, Gonzalez F, Acuna C. 1993. Influence of Layer V of Area 18 of the Cat Visual Cortex on Responses of Cells in Layer V of Area 17 to Stimuli of High Velocity. Exp Brain Res 93: 363-366

Anderson CH, Van Essen DC. 1987. Shifter Circuits: A Computational Strategy for Dynamic Aspects of Visual Processing. Proc Natl Acad Sci U S A 84: 6297-6301

Angelucci A, Bressloff PC. 2006. Contribution of Feedforward, Lateral and Feedback Connections to the Classical Receptive Field Center and Extra-Classical Receptive Field Surround of Primate V1 Neurons. Prog Brain Res Volume 154, Part A: 93-120

Angelucci A, Bullier J. 2003. Reaching Beyond the Classical Receptive Field of V1 Neurons: Horizontal or Feedback Axons? J Physiol Paris 97: 141-154

Angelucci A, Levitt JB, Lund JS. 2002. Anatomical Origins of the Classical Receptive Field and Modulatory Surround Field of Single Neurons in Macaque Visual Cortical Area V1. Prog Brain Res 136: 373-388

Angelucci A, Sainsbury K. 2006. Contribution of Feedforward Thalamic Afferents and Corticogeniculate Feedback to the Spatial Summation Area of Macaque V1 and Lgn. J Comp Neurol 498: 330-351

Arbelaez P, Maire M, Fowlkes C, Malik J. 2011. Contour Detection and Hierarchical Image Segmentation. IEEE Trans Pattern Anal Mach Intell 33: 898-916

Babaud J, Witkin AP, Baudin M, Duda RO. 1986. Uniqueness of the Gaussian Kernel for Scale-Space Filtering. IEEE Trans Pattern Anal Mach Intell 8: 26-33

Baluch F, Itti L. 2010. Training Top-Down Attention Improves Performance on a Triple-Conjunction Search Task. PLoS One 5: e9127

Bar M. 2004. Visual Objects in Context. Nat Rev Neurosci 5: 617-629

Barberini CL, Cohen MR, Wandell BA, Newsome WT. 2005. Cone Signal Interactions in Direction-Selective Neurons in the Middle Temporal Visual Area. J Vis 5: 603-621

Barlow H, Tripathy SP. 1997. Correspondence Noise and Signal Pooling in the Detection of Coherent Visual Motion. J Neurosci 17: 7954-7966

Barnikol UB, Amunts K, Dammers J, Mohlberg H, Fieseler T, et al. 2006. Pattern Reversal Visual Evoked Responses of V1/V2 and V5/Mt as Revealed by Meg Combined with Probabilistic Cytoarchitectonic Maps. Neuroimage 31: 86-108

Battaglini PP, Squatrito S, Galletti C, Maioli MG, Sanseverino Riva E. 1982. Bilateral Projections from the Visual Cortex to the Striatum in the Cat. Exp Brain Res 47: 28-32

Bay H, Ess A, Tuytelaars T, Van Gool L. 2008. Speeded-up Robust Features (SURF). CVIU 110: 346-359

Bayerl P, Neumann H. 2006. Disambiguating Visual Motion by Form-Motion Interaction—a Computational Model. International Journal of Computer Vision 72: 27-45

Beck C, Neumann H. 2010. Interactions of Motion and Form in Visual Cortex - a Neural Model. J Physiol Paris 104: 61-70

Berzhanskaya J, Grossberg S, Mingolla E. 2007. Laminar Cortical Dynamics of Visual Form and Motion Interactions During Coherent Object Motion Perception. Spat Vis 20: 337-395

Biederman I. 1987. Recognition-by-Components: A Theory of Human Image Understanding. Psychol Rev 94: 115-147

Boehler CN, Tsotsos JK, Schoenfeld MA, Heinze HJ, Hopf JM. 2009. The Center-Surround Profile of the Focus of Attention Arises from Recurrent Processing in Visual Cortex. Cereb Cortex 19: 982-991

Borji A, Itti L. 2014. Defending Yarbus: Eye Movements Reveal Observers' Task. J Vis 14: 29

Borji A, Sihite DN, Itti L. 2012. An Object-Based Bayesian Framework for Top-Down Visual Attention. Presented at AAAI, Conference

Born RT, Bradley DC. 2005. Structure and Function of Visual Area MT. Annu Rev Neurosci 28: 157-189

Boynton GM, Hegde J. 2004. Visual Cortex: The Continuing Puzzle of Area V2. Curr Biol 14: R523-524

Brown M, Lowe DG. 2007. Automatic Panoramic Image Stitching Using Invariant Features. International journal of computer vision 74: 59-73

Bruce ND, Tsotsos JK. 2009a. Saliency, Attention, and Visual Search: An Information Theoretic Approach. J Vis 9: 5 1-24

Bruce ND, Tsotsos JK. 2009b. Spatiotemporal Saliency: Towards a Hierarchical Representation of Visual Saliency In Attention in Cognitive Systems, pp. 98-111: Springer

Bullier J. 2001. Integrated Model of Visual Processing. Brain Res Rev 36: 96-107

Bullier J, Hupe JM, James AC, Girard P. 2001. The Role of Feedback Connections in Shaping the Responses of Visual Cortical Neurons. Prog Brain Res 134: 193-204

Bullier J, Kennedy H, Salinger W. 1984. Branching and Laminar Origin of Projections between Visual Cortical Areas in the Cat. J Comp Neurol 228: 329-341

Bushnell MC, Goldberg ME, Robinson DL. 1981. Behavioral Enhancement of Visual Responses in Monkey Cerebral Cortex. I. Modulation in Posterior Parietal Cortex Related to Selective Visual Attention. 755-772 pp.

Canny J. 1986. A Computational Approach to Edge Detection. IEEE Trans Pattern Anal Mach Intell 8: 679-698

Carpenter GA, Grossberg S. 1990. Art 3: Hierarchical Search Using Chemical Transmitters in Self-Organizing Pattern Recognition Architectures. Neural Netw 3: 129-152

Carpenter GA, Grossberg S, Rosen DB. 1991. Art 2-A: An Adaptive Resonance Algorithm for Rapid Category Learning and Recognition. Neural Netw 4: 493-504

Casagrande VA, Xu X. 2004. Parallel Visual Pathways: A Comparative Perspective. Vis Neurosci: 1808

Chelazzi L, Miller EK, Duncan J, Desimone R. 2001. Responses of Neurons in Macaque Area V4 During Memory-Guided Visual Search. Cereb Cortex 11: 761-772

Cheng S-F, Chen W, Sundaram H. 1998. Semantic Visual Templates: Linking Visual Features to Semantics. Presented at ICIP, Conference

Chun MM, Jiang Y. 1998. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. Cogn Psychol 36: 28-71

Cucchiara R, Grana C, Piccardi M, Prati A. 2003. Detecting Moving Objects, Ghosts, and Shadows in Video Streams. IEEE Trans Pattern Anal Mach Intell 25: 1337-1342

Culhane S, Tsotsos J. 1992. An Attentional Prototype for Early Vision. Presented at ECCV, Conference

Dalal N, Triggs B. 2005. Histograms of Oriented Gradients for Human Detection. Presented at CVPR, Conference

Daugman JG. 1980. Two-Dimensional Spectral Analysis of Cortical Receptive Field Profiles. Vision Res 20: 847-856

Davson H. 2012. Physiology of the Eye. Elsevier.

Delon-Martin C, Dojat M, Seghier M, Warnking J, Segebarth C, Bullier J. 2000. Illusory Contours Have Retinotopic Representations in V1. Neuroimage 11: S695

Derrington AM, Krauskopf J, Lennie P. 1984. Chromatic Mechanisms in Lateral Geniculate Nucleus of Macaque. J Physiol 357: 241-265

Desimone R, Duncan J. 1995. Neural Mechanisms of Selective Visual Attention. Annu Rev Neurosci 18: 193-222

Distefano III, J. J., Stubberud AR, Williams IJ. 1967. Feedback and Control Systems. Schaum.

Dobbins A, Zucker SW, Cynader MS. 1987. Endstopped Neurons in the Visual Cortex as a Substrate for Calculating Curvature. Nature 329: 438-441

Duffy CJ. 1998. Mst Neurons Respond to Optic Flow and Translational Movement. J Neurophysiol 80: 1816-1827

Duffy CJ, Wurtz RH. 1991. Sensitivity of Mst Neurons to Optic Flow Stimuli. I. A Continuum of Response Selectivity to Large-Field Stimuli. J Neurophysiol 65: 1329-1345

Duffy CJ, Wurtz RH. 1997. Medial Superior Temporal Area Neurons Respond to Speed Patterns in Optic Flow. The Journal of neuroscience 17: 2839-2851

Einevoll GT, Plesser HE. 2012. Extended Difference-of-Gaussians Model Incorporating Cortical Feedback for Relay Cells in the Lateral Geniculate Nucleus of Cat. Cogn Neurodyn 6: 307-324

Fei-Fei L, Fergus R, Perona P. 2006. One-Shot Learning of Object Categories. IEEE Trans Pattern Anal Mach Intell 28: 594-611

Feldman JA, Ballard DH. 1982. Connectionist Models and Their Properties. Cognitive Science 6: 205-254

Felleman DJ, Lim H, Xiao Y, Wang Y, Eriksson A, Parajuli A. 2015. The Representation of Orientation in Macaque V2: Four Stripes Not Three. Cereb Cortex 25: 2354-2369

Felleman DJ, Van Essen DC. 1991. Distributed Hierarchical Processing in the Primate Cerebral Cortex. Cereb Cortex 1: 1-47

Felzenszwalb PF, Girshick RB, Mcallester D, Ramanan D. 2010. Object Detection with Discriminatively Trained Part-Based Models. IEEE Trans Pattern Anal Mach Intell 32: 1627-1645

Ferrer JM, Kato N, Price DJ. 1992. Organization of Association Projections from Area 17 to Areas 18 and 19 and to Suprasylvian Areas in the Cat's Visual Cortex. J Comp Neurol 316: 261-278

Ferrer JM, Price DJ, Blakemore C. 1988. The Organization of Corticocortical Projections from Area 17 to Area 18 of the Cat's Visual Cortex. Proc R Soc Lond B Biol Sci 233: 77-98

Ffytche DH, Zeki S. 1996. Brain Activity Related to the Perception of Illusory Contours. Neuroimage 3: 104-108

Fidler S, Leonardis A. 2007. Towards Scalable Representations of Object Categories: Learning a Hierarchy of Parts. Presented at CVPR, Conference

Field DJ. 1987. Relations between the Statistics of Natural Images and the Response Properties of Cortical Cells. JOSA A 4: 2379-2394

Fogel I, Sagi D. 1989. Gabor Filters as Texture Discriminator. Biol Cybern 61: 103-113

Foxe JJ, Simpson GV. 2002. Flow of Activation from V1 to Frontal Cortex in Humans. A Framework for Defining "Early" Visual Processing. Exp Brain Res 142: 139-150

Fukushima K. 1980. Neocognitron: A Self Organizing Neural Network Model for a Mechanism of Pattern Recognition Unaffected by Shift in Position. Biol Cybern 36: 193-

202

Fukushima K. 1986. A Neural Network Model for Selective Attention in Visual Pattern Recognition. Biol Cybern 55: 5-15

Fukushima K. 1988. Neocognitron: A Hierarchical Neural Network Capable of Visual Pattern Recognition. Neural Netw 1: 119-130

Fukushima K. 2001. Recognition of Partly Occluded Patterns: A Neural Network Model. Biol Cybern 84: 251-259

Fukushima K. 2005. Restoring Partly Occluded Patterns: A Neural Network Model. Neural Netw 18: 33-43

Fukushima K. 2007. Neocognitron. Scholarpedia 2: 1717

Fukushima K. 2013. Artificial Vision by Multi-Layered Neural Networks: Neocognitron and Its Advances. Neural Netw 37: 103-119

Gao D, Mahadevan V, Vasconcelos N. 2008. On the Plausibility of the Discriminant Center-Surround Hypothesis for Visual Saliency. J Vis 8: 13 11-18

Georgeson MA, May KA, Freeman TC, Hesse GS. 2007. From Filters to Features: Scale–Space Analysis of Edge and Blur Coding in Human Vision. J Vis 7: 7

Ginsburg AP. 1975. Is the Illusory Triangle Physical or Imaginary? Nature

Girard P, Bullier J. 1989. Visual Activity in Area V2 During Reversible Inactivation of Area 17 in the Macaque Monkey. J Neurophysiol 62: 1287-1302

Girard P, Salin PA, Bullier J. 1992. Response Selectivity of Neurons in Area Mt of the Macaque Monkey During Reversible Inactivation of Area V1. J Neurophysiol 67: 1437-1446

Gobbini MI, Haxby JV. 2007. Neural Systems for Recognition of Familiar Faces. Neuropsychologia 45: 32-41

Goebel R, Khorram-Sefat D, Muckli L, Hacker H, Singer W. 1998. The Constructive Nature of Vision: Direct Evidence from Functional Magnetic Resonance Imaging Studies of Apparent Motion and Motion Imagery. Eur J Neurosci 10: 1563-1573

Graziano MS, Andersen RA, Snowden RJ. 1994. Tuning of Mst Neurons to Spiral Motions. J Neurosci 14: 54-67

Greene MR, Liu T, Wolfe JM. 2012. Reconsidering Yarbus: A Failure to Predict Observers' Task from Eye Movement Patterns. Vision research 62: 1-8

Gregoriou GG, Gotts SJ, Zhou H, Desimone R. 2009. High-Frequency, Long-Range Coupling between Prefrontal and Visual Cortex During Attention. Science 324: 1207-1210

Grigorescu C, Petkov N, Westenberg MA. 2003. Contour Detection Based on Nonclassical Receptive Field Inhibition. IEEE Trans Image Process 12: 729-739

Grill-Spector K, Kourtzi Z, Kanwisher N. 2001. The Lateral Occipital Complex and Its Role in Object Recognition. Vision Res 41: 1409-1422

Grossberg S. 1987. Competitive Learning: From Interactive Activation to Adaptive Resonance. Cognitive Science 11: 23-63

Haenny PE, Maunsell JH, Schiller PH. 1988. State Dependent Activity in Monkey Visual Cortex. Ii. Retinal and Extraretinal Factors in V4. Exp Brain Res 69: 245-259

Haji-Abolhassani A, Clark JJ. 2014. An Inverse Yarbus Process: Predicting Observers' Task from Eye Movement Patterns. Vision research 103: 127-142

Han B, Comaniciu D, Zhu Y, Davis LS. 2008. Sequential Kernel Density Approximation and Its Application to Real-Time Visual Tracking. IEEE Trans Pattern Anal Mach Intell 30: 1186-1197

Han B, Davis LS. 2012. Density-Based Multifeature Background Subtraction with Support Vector Machine. IEEE Trans Pattern Anal Mach Intell 34: 1017-1023

Han S, Weaver JA, Murray SO, Kang X, Yund EW, Woods DL. 2002. Hemispheric Asymmetry in Global/Local Processing: Effects of Stimulus Position and Spatial Frequency. Neuroimage 17: 1290-1299

Hendrickson AE, Wilson JR, Ogren MP. 1978. The Neuroanatomical Organization of Pathways between the Dorsal Lateral Geniculate Nucleus and Visual Cortex in Old World and New World Primates. J Comp Neurol 182: 123-136

Henry GH, Salin PA, Bullier J. 1991. Projections from Areas 18 and 19 to Cat Striate Cortex: Divergence and Laminar Specificity. Eur J Neurosci 3: 186-200

Herzog MH, Clarke AM. 2014. Why Vision Is Not Both Hierarchical and Feedforward. Front Comput Neurosci 8

Hirsch J, Delapaz RL, Relkin NR, Victor J, Kim K, et al. 1995. Illusory Contours Activate

Specific Regions in Human Visual Cortex: Evidence from Functional Magnetic Resonance Imaging. Proc Natl Acad Sci U S A 92: 6469-6473

Hubel DH. 1963. Integrative Processes in Central Visual Pathways of the Cat. J Opt Soc Am 53: 58-66

Hubel DH, Wiesel TN. 1959. Receptive Fields of Single Neurones in the Cat's Striate Cortex. J Physiol 148: 574-591

Hubel DH, Wiesel TN. 1962. Receptive Fields, Binocular Interaction and Functional Architecture in the Cat's Visual Cortex. J Physiol 160: 106-154

Hubel DH, Wiesel TN. 1977. Ferrier Lecture. Functional Architecture of Macaque Monkey Visual Cortex. Proc R Soc Lond B Biol Sci 198: 1-59

Humphreys GW, Muller HJ. 1993. Search Via Recursive Rejection (Serr): A Connectionist Model of Visual Search. Cognitive Psychology 25: 43-110

Hupé JM, James AC, Girard P, Lomber SG, Payne BR, Bullier J. 2001. Feedback Connections Act on the Early Part of the Responses in Monkey Visual Cortex. J Neurophysiol 85: 134-145

Hupé JM, James AC, Payne BR, Lomber SG, Girard P, Bullier J. 1998. Cortical Feedback Improves Discrimination between Figure and Background by V1, V2 and V3 Neurons. Nature 394: 784-787

Hyvarinen A. 1999. Fast and Robust Fixed-Point Algorithms for Independent Component Analysis. IEEE Trans Neural Netw 10: 626-634

Irvin GE, Norton TT, Sesma MA, Casagrande VA. 1986. W-Like Response Properties of Interlaminar Zone Cells in the Lateral Geniculate Nucleus of a Primate (Galago Crassicaudatus). Brain Res 362: 254-270

Itti L. 2007. Visual Salience. Scholarpedia 2: 3327

Itti L, Koch C. 2001. Feature Combination Strategies for Saliency-Based Visual Attention Systems. Journal of Electronic Imaging 10: 161-169

Itti L, Koch C, Niebur E. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. IEEE Trans Pattern Anal Mach Intell 20: 1254-1259

Jonas P, Buzsaki G. 2007. Neural Inhibition. Scholarpedia 2: 3286

Jones HE, Wang W, Sillito AM. 2002. Spatial Organization and Magnitude of Orientation

Contrast Interactions in Primate V1. J Neurophysiol 88: 2796-2808

Jones JP, Palmer LA. 1987. An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex. J Neurophysiol 58: 1233-1258

Jones MJ, Sinha P, Vetter T, Poggio T. 1997. Top-Down Learning of Low-Level Vision Tasks. Curr Biol 7: 991-994

Kandel E, Schwartz J, Jessell T. 2013. Principles of Neural Science, Fifth Edition. McGraw-Hill Education.

Kaplan E, Marcus S, So YT. 1979. Effects of Dark Adaptation on Spatial and Temporal Properties of Receptive Fields in Cat Lateral Geniculate Nucleus. J Physiol 294: 561-580

Kaplan E, Shapley RM. 1986. The Primate Retina Contains Two Types of Ganglion Cells, with High and Low Contrast Sensitivity. Proc Natl Acad Sci U S A 83: 2755-2757

Kastner S, Pinsk MA, De Weerd P, Desimone R, Ungerleider LG. 1999. Increased Activity in Human Visual Cortex During Directed Attention in the Absence of Visual Stimulation. Neuron 22: 751-761

Kastner S, Ungerleider LG. 2000. Mechanisms of Visual Attention in the Human Cortex. Annu Rev Neurosci 23: 315-341

Kawano K, Shidara M, Watanabe Y, Yamane S. 1994. Neural Activity in Cortical Area Mst of Alert Monkey During Ocular Following Responses. J Neurophysiol 71: 2305-2324

Keller GB, Bonhoeffer T, Hubener M. 2012. Sensorimotor Mismatch Signals in Primary Visual Cortex of the Behaving Mouse. Neuron 74: 809-815

Kersten D, Mamassian P, Yuille A. 2004. Object Perception as Bayesian Inference. Annu. Rev. Psychol. 55: 271-304

Koch C, Ullman S. 1985. Shifts in Selective Visual Attention: Towards the Underlying Neural Circuitry. Hum Neurobiol 4: 219-227

Koenderink JJ. 1984. The Structure of Images. Biol Cybern 50: 363-370

Kourtzi Z, Dicarlo JJ. 2006. Learning and Neural Plasticity in Visual Object Recognition. Curr Opin Neurobiol 16: 152-158

Kravitz DJ, Saleem KS, Baker CI, Ungerleider LG, Mishkin M. 2013. The Ventral Visual Pathway: An Expanded Neural Framework for the Processing of Object Quality. Trends Cogn Sci 17: 26-49

Kveraga K, Boshyan J, Bar M. 2007. Magnocellular Projections as the Trigger of Top-Down Facilitation in Recognition. J Neurosci 27: 13232-13240

Lamme VA, Roelfsema PR. 2000. The Distinct Modes of Vision Offered by Feedforward and Recurrent Processing. Trends Neurosci 23: 571-579

Lamme VA, Super H, Spekreijse H. 1998. Feedforward, Horizontal, and Feedback Processing in the Visual Cortex. Curr Opin Neurobiol 8: 529-535

Larsson J, Amunts K, Gulyas B, Malikovic A, Zilles K, Roland PE. 1999. Neuronal Correlates of Real and Illusory Contour Perception: Functional Anatomy with Pet. Eur J Neurosci 11: 4024-4036

Lecun BB, Denker JS, Henderson D, Howard RE, Hubbard W, Jackel LD. 1990. Handwritten Digit Recognition with a Back-Propagation Network. Presented at Advances in neural information processing systems, Conference

Lecun Y, Bengio Y, Hinton G. 2015. Deep Learning. Nature 521: 436-444

Lecun Y, Bottou L, Bengio Y, Haffner P. 1998. Gradient-Based Learning Applied to Document Recognition. Proceedings of the IEEE 86: 2278-2324

Lee DS. 2005. Effective Gaussian Mixture Learning for Video Background Subtraction. IEEE Trans Pattern Anal Mach Intell 27: 827-832

Lee TS, Mumford D. 2003. Hierarchical Bayesian Inference in the Visual Cortex. J Opt Soc Am A Opt Image Sci Vis 20: 1434-1448

Lee TS, Nguyen M. 2001. Dynamics of Subjective Contour Formation in the Early Visual Cortex. Proc Natl Acad Sci U S A 98: 1907-1911

Lesher GW, Mingolla E. 1993. The Role of Edges and Line-Ends in Illusory Contour Formation. Vision Res 33: 2253-2270

Levitt JB, Kiper DC, Movshon JA. 1994. Receptive Fields and Functional Architecture of Macaque V2. J Neurophysiol 71: 2517-2542

Li Z. 2002. A Saliency Map in Primary Visual Cortex. Trends Cogn Sci 6: 9-16

Li Z, Snowden RJ. 2006. A Theory of a Saliency Map in Primary Visual Cortex (V1) Tested by Psychophysics of Colour–Orientation Interference in Texture Segmentation. Visual cognition 14: 911-933

Lienhart R, Maydt J. 2002. An Extended Set of Haar-Like Features for Rapid Object

Detection. Presented at ICIP, Conference

Lindeberg T. 1991. Discrete Scale-Space Theory and the Scale-Space Primal Sketch. Royal Institute of Technology

Lindeberg T. 1993. Scale-Space Theory in Computer Vision. Springer.

Lindeberg T. 1994. Scale Selection for Differential Operators. Springer.

Lindeberg T. 1996. Scale-Space: A Framework for Handling Image Structures at Multiple Scales. CERN: 27-38

Lindeberg T. 1998. Feature Detection with Automatic Scale Selection. International journal of computer vision 30: 79-116

Lindeberg T, Florack L. 1994. Foveal Scale-Space and the Linear Increase of Receptive Field Size as a Function of Eccentricity

Lowe DG. 1999. Object Recognition from Local Scale-Invariant Features. Presented at ICCV, Conference

Lowe DG. 2004. Distinctive Image Features from Scale-Invariant Keypoints. International Journal of Computer Vision 60: 91-110

Luck SJ, Chelazzi L, Hillyard SA, Desimone R. 1997. Neural Mechanisms of Spatial Selective Attention in Areas V1, V2, and V4 of Macaque Visual Cortex. J Neurophysiol 77: 24-42

Lueschow A, Miller EK, Desimone R. 1994. Inferior Temporal Mechanisms for Invariant Object Recognition. Cereb Cortex 4: 523-531

Lund JS. 1988. Anatomical Organization of Macaque Monkey Striate Visual Cortex. Annu Rev Neurosci 11: 253-288

Marr D. 1982. Vision: A Computational Investigation into the Human Representation and Processing of Visual Information. Henry Holt and Co., Inc.

Marr D, Hildreth E. 1980. Theory of Edge Detection. Proc R Soc Lond B Biol Sci 207: 187-217

Martin DR, Fowlkes CC, Malik J. 2004. Learning to Detect Natural Image Boundaries Using Local Brightness, Color, and Texture Cues. IEEE Trans Pattern Anal Mach Intell 26: 530-549

Martinez-Conde S, Cudeiro J, Grieve KL, Rodriguez R, Rivadulla C, Acuna C. 1999. Effects of Feedback Projections from Area 18 Layers 2/3 to Area 17 Layers 2/3 in the Cat Visual Cortex. J Neurophysiol 82: 2667-2675

Matthews N, Welch L. 1997. The Effect of Inducer Polarity and Contrast on the Perception of Illusory Figures. Perception 26: 1431-1443

Maunsell JH, Ghose GM, Assad JA, Mcadams CJ, Boudreau CE, Noerager BD. 1999. Visual Response Latencies of Magnocellular and Parvocellular Lgn Neurons in Macaque Monkeys. Vis Neurosci 16: 1-14

Maunsell JH, Gibson JR. 1992. Visual Response Latencies in Striate Cortex of the Macaque Monkey. J Neurophysiol 68: 1332-1344

Maunsell JH, Van Essen DC. 1983. The Connections of the Middle Temporal Visual Area (Mt) and Their Relationship to a Cortical Hierarchy in the Macaque Monkey. J Neurosci 3: 2563-2586

Mehrotra R, Namuduri KR, Ranganathan N. 1992. Gabor Filter-Based Edge Detection. Pattern Recognition 25: 1479-1494

Mendola JD, Dale AM, Fischl B, Liu AK, Tootell RB. 1999. The Representation of Illusory and Real Contours in Human Cortical Visual Areas Revealed by Functional Magnetic Resonance Imaging. J Neurosci 19: 8560-8572

Michelson AA. 1995. Studies in Optics. Dover Publications.

Mishkin M, Ungerleider LG. 1982. Contribution of Striate Inputs to the Visuospatial Functions of Parieto-Preoccipital Cortex in Monkeys. Behav Brain Res 6: 57-77

Mishkin M, Ungerleider LG, Macko KA. 1983. Object Vision and Spatial Vision: Two Cortical Pathways. Trends Neurosci 6: 414-417

Motter BC. 1993. Focal Attention Produces Spatially Selective Processing in Visual Cortical Areas V1, V2, and V4 in the Presence of Competing Stimuli. J Neurophysiol 70: 909-919

Murray MM, Wylie GR, Higgins BA, Javitt DC, Schroeder CE, Foxe JJ. 2002. The Spatiotemporal Dynamics of Illusory Contour Processing: Combined High-Density Electrical Mapping, Source Analysis, and Functional Magnetic Resonance Imaging. J Neurosci 22: 5055-5073

Nelson JI, Frost BJ. 1978. Orientation-Selective Inhibition from Beyond the Classic Visual

Receptive Field. Brain Res 139: 359-365

Norman J. 2002. Two Visual Systems and Two Theories of Perception: An Attempt to Reconcile the Constructivist and Ecological Approaches. Behav Brain Sci 25: 73-96; discussion 96-144

Norton TT, Casagrande VA. 1982. Laminar Organization of Receptive-Field Properties in Lateral Geniculate Nucleus of Bush Baby (Galago Crassicaudatus). 715-741 pp.

Nowak LG, James AC, Bullier J. 1997. Corticocortical Connections between Visual Areas 17 and 18a of the Rat Studied in Vitro: Spatial and Temporal Organisation of Functional Synaptic Responses. Exp Brain Res 117: 219-241

Nowak LG, Munk MH, Girard P, Bullier J. 1995. Visual Latencies in Areas V1 and V2 of the Macaque Monkey. Vis Neurosci 12: 371-384

Oliva A. 2005. Gist of the Scene. Neurobiology of attention 696: 64

Oliva A, Torralba A. 2007. The Role of Context in Object Recognition. Trends Cogn Sci 11: 520-527

Osborne LC, Bialek W, Lisberger SG. 2004. Time Course of Information About Motion Direction in Visual Area Mt of Macaque Monkeys. J Neurosci 24: 3210-3222

Pack CC, Born RT. 2001. Temporal Dynamics of a Neural Solution to the Aperture Problem in Visual Area Mt of Macaque Brain. Nature 409: 1040-1042

Papari G, Campisi P, Petkov N, Neri A. 2007. A Biologically Motivated Multiresolution Approach to Contour Detection. EURASIP 2007: 071828

Pascual-Leone A, Walsh V. 2001. Fast Backprojections from the Motion to the Primary Visual Area Necessary for Visual Awareness. Science 292: 510-512

Pegna AJ, Khateb A, Murray MM, Landis T, Michel CM. 2002. Neural Processing of Illusory and Real Contours Revealed by High-Density Erp Mapping. Neuroreport 13: 965-968

Peichl L, Wassle H. 1979. Size, Scatter and Coverage of Ganglion Cell Receptive Field Centres in the Cat Retina. J Physiol 291: 117-141

Perkel DJ, Bullier J, Kennedy H. 1986. Topography of the Afferent Connectivity of Area 17 in the Macaque Monkey: A Double-Labelling Study. J Comp Neurol 253: 374-402

Perry VH, Oehler R, Cowey A. 1984. Retinal Ganglion Cells That Project to the Dorsal Lateral Geniculate Nucleus in the Macaque Monkey. Neuroscience 12: 1101-1123

Phaf RH, Van Der Heijden AH, Hudson PT. 1990. Slam: A Connectionist Model for Attention in Visual Selection Tasks. Cogn Psychol 22: 273-341

Poggio T, Gamble EB, Little JJ. 1988. Parallel Integration of Vision Modules. Science 242: 436-440

Quiroga RQ, Reddy L, Kreiman G, Koch C, Fried I. 2005. Invariant Visual Representation by Single Neurons in the Human Brain. Nature 435: 1102-1107

Rajimehr R. 2004. Static Motion Aftereffect Does Not Modulate Positional Representations in Early Visual Areas. Cognitive Brain Res 20: 323-327

Ramachandran VS, Ruskin D, Cobb S, Rogers-Ramachandran D, Tyler CW. 1994. On the Perception of Illusory Contours. Vision Res 34: 3145-3152

Ramsden BM, Hung CP, Roe AW. 2001. Real and Illusory Contour Processing in Area V1 of the Primate: A Cortical Balancing Act. Cereb Cortex 11: 648-665

Rauss K, Schwartz S, Pourtois G. 2011. Top-Down Effects on Early Visual Processing in Humans: A Predictive Coding Framework. Neurosci Biobehav Rev 35: 1237-1253

Ren X. 2008. Multi-Scale Improves Boundary Detection in Natural Images. Presented at ECCV, Conference

Reynolds JH, Chelazzi L, Desimone R. 1999. Competitive Mechanisms Subserve Attention in Macaque Areas V2 and V4. J Neurosci 19: 1736-1753

Riesenhuber M, Poggio T. 1999. Hierarchical Models of Object Recognition in Cortex. Nat Neurosci 2: 1019-1025

Rodieck RW. 1965. Quantitative Analysis of Cat Retinal Ganglion Cell Response to Visual Stimuli. Vision Res 5: 583-601

Rodriguez-Sanchez AJ, Tsotsos JK. 2011. The Importance of Intermediate Representations for the Modeling of 2d Shape Detection: Endstopping and Curvature Tuned Computations. Presented at CVPR, Conference

Rodriguez-Sanchez AJ, Tsotsos JK. 2012. The Roles of Endstopped and Curvature Tuned Computations in a Hierarchical Representation of 2d Shape. PLoS One 7: e42058

Roelfsema PR. 2006. Cortical Algorithms for Perceptual Grouping. Annu Rev Neurosci 29: 203-227

Rolls ET. 2000. Functions of the Primate Temporal Lobe Cortical Visual Areas in Invariant

Visual Object and Face Recognition. Neuron 27: 205-218

Rust NC, Mante V, Simoncelli EP, Movshon JA. 2006. How Mt Cells Analyze the Motion of Visual Patterns. Nat Neurosci 9: 1421-1431

Salin PA, Bullier J. 1995. Corticocortical Connections in the Visual System: Structure and Function. Physiol Rev 75: 107-154

Salin PA, Girard P, Kennedy H, Bullier J. 1992. Visuotopic Organization of Corticocortical Connections in the Visual System of the Cat. J Comp Neurol 320: 415-434

Sandell JH, Schiller PH. 1982. Effect of Cooling Area 18 on Striate Cortex Cells in the Squirrel Monkey. J Neurophysiol 48: 38-48

Schiller PH, Finlay BL, Volman SF. 1976. Quantitative Studies of Single-Cell Properties in Monkey Striate Cortex. Iii. Spatial Frequency. J Neurophysiol 39: 1334-1351

Schmolesky MT, Wang Y, Hanes DP, Thompson KG, Leutgeb S, et al. 1998. Signal Timing across the Macaque Visual System. J Neurophysiol 79: 3272-3278

Seghier M, Dojat M, Delon-Martin C, Rubin C, Warnking J, et al. 2000a. Moving Illusory Contours Activate Primary Visual Cortex: An Fmri Study. Cereb Cortex 10: 663-670

Seghier M, Dojat M, Delon-Martin C, Rubin C, Warnking J, et al. 2000b. Moving Illusory Contours Activate Primary Visual Cortex: An Fmri Study. Cerebral Cortex 10: 663-670

Seghier ML, Vuilleumier P. 2006. Functional Neuroimaging Findings on the Human Perception of Illusory Contours. Neurosci Biobehav Rev 30: 595-612

Serre T, Riesenhuber M. 2004. Realistic Modeling of Simple and Complex Cell Tuning in the Hmax Model, and Implications for Invariant Object Recognition in Cortex. Technical Reports Rep. MIT-CSAIL-TR-2004-052, Massachusetts Institute of Technology Cambridge

Serre T, Wolf L, Bileschi S, Riesenhuber M, Poggio T. 2007. Robust Object Recognition with Cortex-Like Mechanisms. IEEE Trans Pattern Anal Mach Intell 29: 411-426

Shipley TF, Kellman PJ. 1992. Strength of Visual Interpolation Depends on the Ratio of Physically Specified to Total Edge Length. Percept Psychophys 52: 97-106

Siagian C, Itti L. 2007. Rapid Biologically-Inspired Scene Classification Using Features Shared with Visual Attention. IEEE Trans Pattern Anal Mach Intell 29: 300-312

Silvanto J, Cowey A, Lavie N, Walsh V. 2005. Striate Cortex (V1) Activity Gates Awareness of Motion. Nat Neurosci 8: 143-144

Simoncelli EP, Heeger DJ. 1998. A Model of Neuronal Responses in Visual Area Mt. Vision Res 38: 743-761

Sincich LC, Park KF, Wohlgemuth MJ, Horton JC. 2004. Bypassing V1: A Direct Geniculate Input to Area Mt. Nat Neurosci 7: 1123-1128

Skrypnyk I, Lowe DG. 2004. Scene Modelling, Recognition and Tracking with Invariant Image Features. Presented at ISMAR, Conference

Smith AT, Wall MB, Williams AL, Singh KD. 2006. Sensitivity to Optic Flow in Human Cortical Areas Mt and Mst. Eur J Neurosci 23: 561-569

Snowden RJ, Treue S, Erickson RG, Andersen RA. 1991. The Response of Area Mt and V1 Neurons to Transparent Motion. J Neurosci 11: 2768-2785

Thune M, Olstad B, Thune N. 1997. Edge Detection in Noisy Data Using Finite Mixture Distribution Analysis. Pattern Recognition 30: 685-699

Tootell RB, Silverman MS, Hamilton SL, Switkes E, De Valois RL. 1988. Functional Anatomy of Macaque Striate Cortex. V. Spatial Frequency. J Neurosci 8: 1610-1624

Torralba A. 2003. Contextual Priming for Object Detection. International Journal of Computer Vision 53: 169-191

Torralba A, Murphy KP, Freeman WT, Rubin MA. ICCV2003: 273-280 vol.271.

Treisman AM, Gelade G. 1980a. A Feature-Integration Theory of Attention. Cogn Psychol 12: 97-136

Treisman AM, Gelade G. 1980b. A Feature-Integration Theory of Attention. Cognitive psychology 12: 97-136

Treue S. 2003. Visual Attention: The Where, What, How and Why of Saliency. Curr Opin Neurobiol 13: 428-432

Treue S, Andersen RA. 1996. Neural Responses to Velocity Gradients in Macaque Cortical Area Mt. Vis Neurosci 13: 797-804

Ts'o DY, Frostig RD, Lieke EE, Grinvald A. 1990. Functional Organization of Primate Visual Cortex Revealed by High Resolution Optical Imaging. Science 249: 417-420

Tsotsos JK. 1990. Analyzing Vision at the Complexity Level. Behav. Brain Sci 13: 423-445

Tsotsos JK. 2011. A Computational Perspective on Visual Attention. MIT Press.

Tsotsos JK, Culhane SM, Kei Wai WY, Lai Y, Davis N, Nuflo F. 1995. Modeling Visual Attention Via Selective Tuning. Artificial Intelligence 78: 507-545

Tsotsos JK, Liu Y, Martinez-Trujillo JC, Pomplun M, Simine E, Zhou K. 2005. Attending to Visual Motion. CVIU 100: 3-40

Tsotsos JK, Rodriguez-Sanchez AJ, Rothenstein AL, Simine E. 2008. The Different Stages of Visual Recognition Need Different Attentional Binding Strategies. Brain Res 1225: 119-132

Ungerleider LG, Desimone R. 1986. Cortical Connections of Visual Area Mt in the Macaque. J Comp Neurol 248: 190-222

Ungerleider LG, Galkin TW, Desimone R, Gattass R. 2008. Cortical Connections of Area V4 in the Macaque. Cereb Cortex 18: 477-499

Van Den Stock J, Tamietto M, Sorger B, Pichon S, Grezes J, De Gelder B. 2011. Cortico-Subcortical Visual, Somatosensory, and Motor Activations for Perceiving Dynamic Whole-Body Emotional Expressions with and without Striate Cortex (V1). Proc Natl Acad Sci U S A 108: 16188-16193

Van Essen DC, Maunsell JHR. 1983. Hierarchical Organization and Functional Streams in the Visual Cortex. Trends Neurosci 6: 370-375

Van Essen DC, Newsome WT, Maunsell JH, Bixby JL. 1986. The Projections from Striate Cortex (V1) to Areas V2 and V3 in the Macaque Monkey: Asymmetries, Areal Boundaries, and Patchy Connections. J Comp Neurol 244: 451-480

Van Laere J. 1993. Vesalius and the Nervous System. Verh K Acad Geneeskd Belg 55: 533-576

Vanrullen R. 2003. Visual Saliency and Spike Timing in the Ventral Visual Pathway. J Physiol Paris 97: 365-377

Vedaldi A, Fulkerson B. ACM Multimedia Firenze, Italy, 2010: 1469-1472. 1874249: ACM.

Vuilleumier P, Valenza N, Landis T. 2001. Explicit and Implicit Perception of Illusory Contours in Unilateral Spatial Neglect: Behavioural and Anatomical Correlates of Preattentive Grouping Mechanisms. Neuropsychologia 39: 597-610

Wade AD, Nelson AJ, Garvin GJ. 2011. A Synthetic Radiological Study of Brain Treatment in Ancient Egyptian Mummies. Homo 62: 248-269

Watanabe T. 1998. High-Level Motion Processing: Computational, Neurobiological, and Psychophysical Perspectives. MIT Press.

Weller RE, Wall JT, Kaas JH. 1984. Cortical Connections of the Middle Temporal Visual Area (Mt) and the Superior Temporal Cortex in Owl Monkeys. J Comp Neurol 228: 81-104

Wilkins RH. 1992. Neurosurgical Classics. Thieme Medical Publishers, Incorporated.

Witkin AP. 1983. Scale-Space Filtering. Presented at IJCAI, Conference

Xu X, Ichida JM, Allison JD, Boyd JD, Bonds AB, Casagrande VA. 2001. A Comparison of Koniocellular, Magnocellular and Parvocellular Receptive Field Properties in the Lateral Geniculate Nucleus of the Owl Monkey. J Physiol 531: 203-218

Yarbus AL, Haigh B, Rigss LA. 1967. Eye Movements and Vision. Plenum press New York.

Zaharescu A, Rothenstein A, Tsotsos JK 2005. Towards a Biologically Plausible Active Visual Search Model In Attention and Performance in Computational Vision, ed. L Paletta, J Tsotsos, E Rome, G Humphreys, pp. 133-147: Springer Berlin / Heidelberg

Zeki S. 1996. Brain Activity Related to the Perception of Illusory Contours. Neuroimage 3: 104-108

Zeng C, Li Y, Li C. 2011. Center–Surround Interaction with Adaptive Inhibition: A Computational Model for Contour Detection. Neuroimage 55: 49-66

Zhang L, Tong MH, Marks TK, Shan H, Cottrell GW. 2008. SUN: A Bayesian Framework for Saliency Using Natural Statistics. J Vis 8: 32 31-20