Universitat Autònoma de Barcelona

Understanding Eye Movements: Psychophysics and a Model of Primary Visual Cortex

David Berga Garreta

Universitat Autònoma de Barcelona

Understanding Eye Movements: Psychophysics and a Model of Primary Visual Cortex

A dissertation submitted by **David Berga Garreta** at Universitat Autònoma de Barcelona to fulfil the degree of **Doctor of Philosophy**.

Bellaterra, June 7, 2019

Director	**Xavier Otazu** Dept. de Ciències de la Computació, Universitat Autònoma de Barcelona Centre de Visió per Computador
Thesis committee	**Dr. Joost van de Weijer** Centre de Visió per Computador **Dr. Zhaoping Li** University of Tuebingen Max Planck Institute for Biological Cybernetics **Dr. Naila Murray** Naver Labs Europe
International evaluators	**Dr. Olivier Penacchio** University of St Andrews **Dr. Seyedarash Akbarinia** Justus Liebig University Giessen

This document was typeset by the author using $\LaTeX\,2_{\varepsilon}$.

"The real hero is always a hero by mistake;
he dreams of being an honest coward like everybody else."
— Umberto Eco, *Travels in Hyperreality*

"The true formula of atheism is not God is dead
even by basing the origin of the function of the father upon his murder,
Freud protects the father – the true formula of atheism is God is unconscious"
—Jacques Lacan, *The Four Fundamental Concepts of Psychoanalysis*

Mary: It's circular. You exist to continue your existence. What's the point?
John: What's the point of your existence?
Mary: To feel. 'Cause you've never done it, you can never know it. But it's as vital as breath. And without it – without love, without anger, without sorrow – breath is just a clock ticking.
— Equilibrium (2002)

Dedicat als meus pares, Javier i Elisabet

Acknowledgements

I would like to thank all the people who supported me during this journey, you all deserve it. In acquiring knowledge the means and the ends must match to discern congruity and understand balance in things. I believe this can only be built through both intro- and extrospection, requiring time, patience, effort, motivation and determination. Considering these are in order, we see that before the ends there is "motivation" in the equation, here is where you take part.

My parents, Javier and Elisabet, have been the ones who motivated me the most. For these dinners, these jokes, these barbecues, these table talks, these vacation seasons, but also your empathy, your humility, your caring, and most importantly, your love. Although, as we all know, "deadlines" killed some time I could spend with you, I have been missing it. I specially dedicate it to you, for being the first and last who I see before taking a flight and the ones when I come back. You and only you could be whom supported me all the way losing myself in this nerdy field. For this and all, I appreciate all your years of sacrifice, you taught me the true and only values in life. I feel proud to have you. To my older brother, Xavier. Only through your patience for letting me break the limits of Windows, getting some clues how this box works. For your modding sessions with Pali (I also include you in the equation) and your past (and current) gaming sessions, these have always been priceless, and must endure. You are the one I mirrored, that is why I am now and here. Thanking to my beloved family, Margarita, Rosa Maria, Carmen, Santi and many more I could not fit in here. For those who also were there, Cinto, Maria, Vicenç, my memories won't die, this is also for you with all my love.

For all my friends. From those I see the most, whom happen to be musicians and that make everything funny and somewhat artisticly meaningful, Alan, Alberto, Artur, Eugeni, Victor, Oscar, Quim, Nico, Guillem, Diego, Arnau... Also to whom had best moments, Eli, Guiu, Bern, Javi, Riky... To the "Pics", Sasu, Yurx, Gina, Carlota, Carmen, and of course, Pic. Also to my current and past flatmates, Rosie, Mara, Christian, Stephane, Kirstie, Eze, Gerard, Kathrin, Jack and David, with whom suffered my experiments with the retropie. Also to my CVC friends, which I shared some funny talks inside or outside the lab, Manu, Sounak, Dena, Pau, Arnau, Andrés, Ali, Amir, Yashing, Marc (Massana), Marc (Paz), Xim, Nilai, Jacopo, Gianmarco, Bamba, Sara, Marta, Lidia, Claire and many more I see every day. Also to the new ones I met from UB, in Galicia and Canada, Toni, Víctor, David, Gela, Markus, Calden, Maneiro, Amir, Bikram, Pariah and Sang Ah.

I know took risks by specifying the order, let's say ordinality was dependent on time, then make your own guess.

I have added another page reserved for the science ones, those who created the base of knowledge for my PhD, the "jedi masters", it was because of them I wrote this thesis.

First of all, I would like to thank my supervisor *Xavier Otazu Porter*, whom supported and helped me with all the experiments and the whole process of the PhD, he was there every time I needed. Only through his pickiness and persistence, his demonstrative and bright explanations and his wide knowledge in various fields, I could make this thesis as it is. I got entangled by his way of thinking about science and engineering while programming with him, with expressions like "patatim-patatam" or "fer el friki". You would know him acknowledging that "XOP" appears in the tricky lines of code. I can reaffirm he really likes to code, to hack the software and to exploit the computer resources. Also to *Carlos Alejandro Párraga*, whom I had been changing funny talks while having a quite "sciency" perspective about them. To my previous UPF tutors *Joan Mora* and *Pascal Landry*, people from the SPECS lab and *Narcís Parés*, who introduced me to the world of science, giving me the opportunity and interest to start the research career. Finally, all my gratitude for the great supervisors I had during my research stays *Xosé Ramón Fernández*, *Xosé Manuel Pardo* and *John Tsotsos*. In so short time I learned a lot and had really nice lessons from you.

Let's see what all these Profs. taught me to write... enjoy.

Abstract

Humans move their eyes in order to learn visual representations of the world. These eye movements depend on distinct factors, either by the scene that we perceive or by our own decisions. To select what is relevant to attend is part of our survival mechanisms and the way we build reality, as we constantly react both consciously and unconsciously to all the stimuli that is projected into our eyes. In this thesis we try to explain (1) how we move our eyes, (2) how to build machines that understand visual information and deploy eye movements, and (3) how to make these machines understand tasks in order to decide for eye movements.

(1) We provided the analysis of eye movement behavior elicited by low-level feature distinctiveness with a dataset of 230 synthetically-generated image patterns. A total of 15 types of stimuli has been generated (e.g. orientation, brightness, color, size, etc.), with 7 feature contrasts for each feature category. Eye-tracking data was collected from 34 participants during the viewing of the dataset, using Free-Viewing and Visual Search task instructions. Results showed that saliency is predominantly and distinctively influenced by: 1. feature type, 2. feature contrast, 3. temporality of fixations, 4. task difficulty and 5. center bias. From such dataset (SID4VAM), we have computed a benchmark of saliency models by testing performance using psychophysical patterns. Model performance has been evaluated considering model inspiration and consistency with human psychophysics. Our study reveals that state-of-the-art Deep Learning saliency models do not perform well with synthetic pattern images, instead, models with Spectral/Fourier inspiration outperform others in saliency metrics and are more consistent with human psychophysical experimentation.

(2) Computations in the primary visual cortex (area V1 or striate cortex) have long been hypothesized to be responsible, among several visual processing mechanisms, of bottom-up visual attention (also named saliency). In order to validate this hypothesis, images from eye tracking datasets have been processed with a biologically plausible model of V1 (named Neurodynamic Saliency Wavelet Model or NSWAM). Following Li's neurodynamic model, we define V1's lateral connections with a network of firing rate neurons, sensitive to visual features such as brightness, color, orientation and scale. Early subcortical processes (i.e. retinal and thalamic) are functionally simulated. The resulting saliency maps are generated from the model output, representing the neuronal activity of V1 projections towards brain areas involved in eye movement control. We want to pinpoint that our unified computational architecture is able to reproduce several visual processes (i.e. brightness,

chromatic induction and visual discomfort) without applying any type of training or optimization and keeping the same parametrization. The model has been extended (NSWAM-CM) with an implementation of the cortical magnification function to define the retinotopical projections towards V1, processing neuronal activity for each distinct view during scene observation. Novel computational definitions of top-down inhibition (in terms of inhibition of return and selection mechanisms), are also proposed to predict attention in Free-Viewing and Visual Search conditions. Results show that our model outpeforms other biologically-inpired models of saliency prediction as well as to predict visual saccade sequences, specifically for nature and synthetic images. We also show how temporal and spatial characteristics of inhibition of return can improve prediction of saccades, as well as how distinct search strategies (in terms of feature-selective or category-specific inhibition) predict attention at distinct image contexts.

(3) Although previous scanpath models have been able to efficiently predict saccades during Free-Viewing, it is well known that stimulus and task instructions can strongly affect eye movement patterns. In particular, task priming has been shown to be crucial to the deployment of eye movements, involving interactions between brain areas related to goal-directed behavior, working and long-term memory in combination with stimulus-driven eye movement neuronal correlates. In our latest study we proposed an extension of the Selective Tuning Attentive Reference Fixation Controller Model based on task demands (STAR-FCT), describing novel computational definitions of Long-Term Memory, Visual Task Executive and Task Working Memory. With these modules we are able to use textual instructions in order to guide the model to attend to specific categories of objects and/or places in the scene. We have designed our memory model by processing a visual hierarchy of low- and high-level features. The relationship between the executive task instructions and the memory representations has been specified using a tree of semantic similarities between the learned features and the object category labels. Results reveal that by using this model, the resulting object localization maps and predicted saccades have a higher probability to fall inside the salient regions depending on the distinct task instructions compared to saliency.

Key words: *saliency, eye movements, attention, visual cortex, horizontal connections, visual search, free-viewing, psychophysics, firing rate, neural networks*

Resum

Els éssers humans mouen els ulls per tal d'aprendre representacions del món. Aquests moviments coulars depenen de diferents factors, tant per la escena que percebem com per decisions pròpies. Seleccionar allò que és rellevant atendre forma part dels nostres mecanismes de supervivència i la manera de construir la realitat, ja que constantment reaccionem tant conscient com inconscientment a tots els estímuls que es projecten als nostres ulls. En aquesta tesi intentaré explicar (1) com movem els ulls, (2) com fer màquines que entenguin la informació visual i executar moviments oculars, i (3) com fer que aquestes màquines entenguin tasques per tal de decidir per aquets moviments oculars.

(1) Hem analitzat del comportament dels moviments oculars provocat per les diferències de característiques de baix nivell amb una base de dades d'imatges composada per 230 patrons generats sintèticament. S'han generat un total de 15 tipus d'estímuls (p.e. orientació, brillantor, color, tamany, etc.), amb 7 contrastos per cada categoría de característica. Les dades de 34 participants s'han pogut col·leccionar a partir d'un seguidor ocular durant la visualització de la base de dades, amb les tasques d'Observació Lliure i Cerca Visual. Els resultats han mostrat que la saliency és predominantment i distinctivament influenciada per: 1. el tipus de característica, 2. el contrast de característiques, 3. la temporalitat de les fixacions, 4. la dificultat de la tasca i 5. l'esbiaixament central. A partir d'aquesta base de dades (SID4VAM) hem computat una comparació dels models de saliency testejant el seu rendiment utilitzant patrons psicofísics. El rendiment dels models s'ha evaluat detallant la influència de la inspiració i la consistència amb els resultats de la psicofísica. El nostre estudi revela que els models en l'estat de l'art en saliency basats Deep Learning no tenen bon rendiment amb patrons sintètics, contràriament, els models d'inspiració Espectral/Fourier en superen el rendiment i són més consistents amb la experimentació psicofísica.

(2) Les computacions de l'escorça visual primària (area V1 o escorça estriada) s'han hipotetitzat com a responsables, entre altres mecanismes de processament visual, de l'atenció visual bottom-up (o també anomenada saliency). Per tal de validar aquesta hipòtesi, s'han processat diferents bades de dades d'imatges amb seguidor ocular a partir d'un model biològicament plausible de V1 (anomenat Neurodyamic Saliency Wavelet Model o NSWAM). Seguint el model neurodinàmic de Li, hem definit les connexions laterals de V1 amb una xarxa de neurones firing rate, sensitives a característiques visuals com la brillantor, el color, la orientació i la escala. Els processos subcorticals inferiors (i.e. retinals i talàmics) s'han mo-

delitzat funcionalment. Els mapes de saliency resultats s'han generat a partir de la sortida del model, representant l'activitat neuronal de V1 cap a les arees del cervell involucrades en el control dels moviments oculars. Fa falta destacar que la nostra arquitectura unificada és capaç de reproduir diferents processos de la visió (i.e. inducció de brillantor, cromàtica i malestar visual) sense aplicar cap tipus d'entrenament ni optimització i seguint la mateixa parametrització. S'ha extès el model (NSWAM-CM) incluint una implementació de la magnificació cortical per tal de definir les projeccions retinotòpiques cap a V1 per cada visualització de la escena. També s'ha proposat la inhibició top-down (en termes d'inhibició de retorn i mecanismes de selecció) per tal de predir l'atenció tant en Observació Lliure com Cerca Visual. Els resultats han demostrat que el model supera en rendiment a altres models biològicament inspirats per a la predicció de saliency i sequències de saccades, en concret en imatges de sintètiques i de natura. Mostrem també com les característiques espaials i temporals de la inhibició de retorn poden millorar la predicció de les saccades, i també les diferents estratègies de cerca (en termes de inhibició selectiva de característica o de categoría) per predir la atenció en contextos diferents.

(3) Tot i que els models de scanpath anteriors han demostrat eficaçment la predicció de saccades en Observació Lliure, cal destacar que tant l'estímul com les instruccions de la tasca poden afectar notablement els patrons de moviment ocular. En particular, el priming de tasca és crucial per a la execució de moviments oculars, involucrant interaccions entre arees cerebrals relacionades amb la conducta orientada a la meta, memòria de treball i de llarg termini en combinació amb les zones neuronals responsables de processar els estímuls. En l'últim estudi, hem proposat d'extendre el Selective Tuning Reference Fixation Controller Model, basat en instruccions de tasca (STAR-FCT), describint noves definicions computacionals de la Memòria de Llarg Termini, l'Executiu de Tasques Visuals i la Memòria de Treball per a la Tasca. A partir d'aquests mòduls hem sigut capaços d'utilitzar instruccions textuals per tal de guiar el model a dirigir la atenció a categoríes específiques d'objecte i/o llocs concrets de la escena. Hem disenyat el nostre model de memòria a partir de una jerarquía de característiques tant d'alt com de baix nivell. La relació entre les instruccions executives de la tasca i les representacions de la memòria s'han especificat utilitzant un arbre de similaritats semàntiques entre les característiques apreses i les anotacions de categoría d'objecte. Els resultats en comparació amb la saliency han mostrat que utilitzant aquest model, tant els mapes de localització d'objecte com les prediccions de saccades tenen major probabilitat de caure en les regions salients depenent de les instruccions.

Paraules clau: *saliency, moviments oculars, atenció, escorça visual, connexions* *horitzontals, cerca visual, observació lliure, psicofísica, firing rate, xarxes neuronals*

Resumen

Los seres humanos movemos los ojos para tal de aprender las representaciones del mundo. Estos movimientos oculares dependen de diferentes factores, tanto de la escena que percibimos como por decisiones propias. Seleccionar todo aquello que es relevante a atender forma parte de nuestros mecanismos de supervivencia y nuestra manera de construir la realidad, ya que constantemente reaccionamos tanto consciente como inconscientemente a todos los estímulos que se proyectan a nuestros ojos. En esta tesis intentaré explicar (1) cómo movemos los ojos, (2) cómo hacer máquinas que entiendan la información visual y ejectuar los movimientos oculares, y (3) cómo hacer que estas sean capaces de entender tareas para tal de decidir por estos movimientos oculares.

(1) Hemos analizado del comportamiento de los movimientos oculares provocado por las diferencias de características de bajo nivel con una base de datos de 230 patrones generados sintéticamente. Se han generado un total de 15 tipos de estímulo (p.e. orientación, brillo, color, tamaño, etc.), con 7 contrastes por cada categoría de característica. Se obtuvieron los datos de 34 participantes a partir de un seguidor ocular durante la visualización de la base de datos, con las tareas de Observación Libre y Búsqueda Visual. Los resultados han mostrado que la saliency es predominante y distintamente influenciada por: 1. el tipo de característica, 2. el contraste de las características, 3. la temporalidad de las fijaciones, 4. la dificultad de la tarea y 5. el sesgo central. A partir de esta base de datos (SID4VAM), hemos computado una compración de los modelos de saliency testeando su rendimiendo utilizando patrones psicofísicos. El rendimiento de los modelos se ha evaluado detallando la influencia de su inspiración y la consistencia con los resultados de la psicofísica. Nuestro estudio revela que los modelos del estado del arte en saliency basados en Deep Learning no tienen buen rendimiento con patrones sintéticos, contrariamente, los modelos de inspiración Espectral/Fourier superan en rendimiento y són más consistentes con la experimentación psicofísica.

(2) Las computaciones de la corteza visual primaria (area V1 o corteza estriada) se hipotetizaron como responsables, entre otros mecanismos de procesamiento visual, de la atención visual bottom-up (también nombrada saliency o saliencia). Para tal de validar esta hipótesis, se han procesado diferentes bases de datos de imágenes con seguidor ocular a partir de un modelo biológicamente plausible de V1 (nombrado Neurodynamic Saliency Wavelet Model o NSWAM). Siguiendo el modelo de Li, hemos definido las conexiones laterales de V1 con una red de neuronas firing rate, sensitivas a características visuales como el brillo, el color, la orientación

y la escala. Los procesos subcorticales inferiores (i.e. retinales y talàmicos) se han modelizado funcionalmente. Los mapas de saliency resultates se generaron a partir de la salida del modelo, representando la actividad neuronal de V1 hacia los correlatos cerebrales involucrados en el control de los movimientos oculares. Hace falta destacar que nuestra arquitectura unificada es capaz de reproduir diferentes processos de la visión (i.e. inducción de brillo, cromática y malestar visual) sin aplicar ningun tipo de entrenamiento ni optimización y siguiendo la misma parametrización. Se ha extendido el modelo (NSWAM-CM) incluyendo una implementación de la magnificación cortical para definir las proyecciones retinotópicas hacia V1 dada cada visualización de la escena. También se ha propuesto definir la inhibición top-down (en términos de la inhibición de retorno y mecanismos de selección) para tal de predecir la atención tanto en Observación Libre como en Búsqueda Visual. Los resultados han mostrado que el modelo supera en rendimiento a otros modelos biológicamente inspirados para la predicción de saliency y secuencias de sacadas, en concreto con imágenes sintéticas y de naturaleza. Hemos mostrado también como las características espaciales y temporales de la inhibición de retorno pueden mejorar la predicción de las sacadas, y también las diferentes estrategias de selección (en términos de inhibición selectiva de característica o de categoría) para predecir la atención en diferentes contextos.

(3) Aunque los modelos de scanpath anteriores han demostrado predecir eficazmente las sacadas en Observación Libre, hace falta destacar que tanto el estímulo como las instrucciones de la tarea pueden afectar notablemente los patrones de movimiento ocular. En particular el primado de tarea es crucial para la ejecución de movimientos oculares, involucrando interacciones entre areas cerebrales relacionadas con la conducta orientada a la meta, la memoria de trabajo y de largo plazo en combinación con los correlatos neuronales responsables de procesar los estímulos. En el último estudio, hemos propuesto extender el Selective Tuning Reference Fixation Controller Model, basado en instrucciones de tarea (STAR-FCT), describiendo nuevas definiciones computacionales de la Memoria a Largo Plazo, el Ejecutivo de Tareas Visuales y la Memoria de Trabajo para la Tarea. A partir de estos módulos hemos sido capaces de utilizar instrucciones textuales para tal de guiar el modelo a dirigir la atención en categorías específicas de objeto y/o zonas concretas de la escena. Hemos diseñado nuestro modelo de memoria a partir de una jerarquía de características tanto de alto como bajo nivel. La relación entre las instrucciones ejecutivas de la tarea y las representaciones de la memoria se han especificado utilizando un árbol de similaridades semánticas entre las características aprendidas y las anotaciones de categoría de objeto. Los resultados en comparación con la saliency han demostrado que utilizando este modelo, tanto los mapas de localización de objeto como las predicciones de scadas tienen mayor probabilidad

de caer en las regiones salientes dependiendo de las instrucciones.

Palabras clave: *saliency, movimentos oculares, atenció, corteza visual, conexiones horitzontales, búsqueda visual, observación libre, psicofísica, firing rate, redes neuronales*

Contents

List of Figures

List of Tables

1 Introduction

1.1 Visual Perception and the Brain

Human perception could be defined as the interpretation of the world through human senses. In these terms, one can say that what we see (perception) is not the same as what there is (physically), as we are limited by our nervous system. It leads to some visual effects coined with the term "illusion" (see Figure 1.1 and [284]). We are able to perceive and recognize from simple shapes (lines, squares, circles...) to complex shapes (objects, faces...) and they can only be perceived in particular conditions of luminance and spectral range. Their characterization describe our feature descriptions of the world, which is done by neuronal computations in our brain. Modern Gestalt psychology [158, 349] tried to explain these phenomena (by grouping visual descriptions), describing several perceptual principles or laws: Proximity, Similarity, Closure, Symmetry, Continuity and Common Fate (see Chapter 2). These laws establish that spatial organization of visual elements in the scene can lead to different percepts.

One of the main tasks that the human visual system has to solve is to detect and identify visual objects in a scene. Object discriminability and recognition can depend on numerous factors, such as scene illumination, how light is reflected to every single object in the scene (determining distinct perceived luminance and chromaticity for each object) and how objects are located in the scene. Human percepts can be measured with psychophysical experimentation [90], where human decisions and behavior (sensation magnitude) are compared to the physical properties of the viewed stimuli (stimulus intensity). Accounting for these effects there is the case of brightness induction, where changes in perceived brightness of a visual target are due to the luminance of its surrounding area. From this statement, we can perceive a visual target and the surrounding area with similar/equal brightness (assimilation) or different (contrast). These brightness differences can induce a change in the perceived brightness of the central area (Fig. 1.1**A**) where two grey patches are perceived distinctively whilst being with same brightness. Similarly, the HVS perceives the chromatic properties of a visual target depending on the chromaticities of its surrounding area. This phenomena is named chromatic induction. This effect is observable on Fig. 1.1**B**, where the central ring from the reference stimulus (left) appears to be “greener" than the central ring from the test (right), which appears to be “bluer". Besides these effects, it is known that specific visual patterns (Fig. 1.1**C**) can cause discomfort, malaise, nausea or even migraine [174, 240]. The spatial properties of visual elements that compose the scene (whether are dense or sparse) and their relative contrast energy (due to its orientation, luminance, chromatic and spatial frequency distributions) can generate hyperexcitability in V1, a possible cause of visual discomfort for certain images.

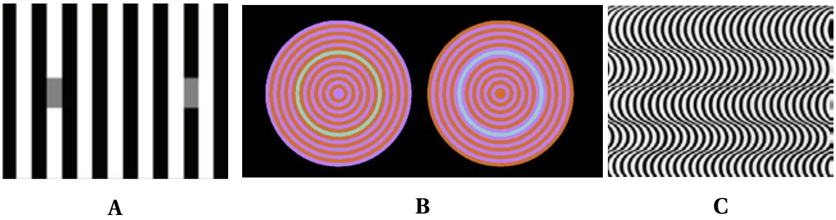


Figure 1.1 – **(A)** Brightness induction from the White effect [353]. The left/right grey patch is surrounded by black/white vertical stripes, which induces to perceive a darker/brighter square patch. **(B)** Example of chromatic induction from Monnier & Shevell’s concentric ring stimuli [207]. Both the left and right central rings are the same color surrounded by red and purple rings. The left ring, in contact with a red inducer, is perceived as greener, while the right ring, in contact with a purple inducer is perceived as bluer. **(C)** Discomfortable image, credit by Nicholas Wade [340].

1.1.1 Main Theories in Visual Attention

Human vision has evolved in order to be more ecologically efficient in our living environments (our ancestors survived in nature environments). In other words, the brain adapts to the environment accounting for its limited processing capacity. The efficient coding and information theory explain this issue as if the visual system discriminates or discards redundant information [15, 225, 290, 384]. Conversely, relevant information is filtered or selected in order to be later processed by higher areas in the brain. In that regard, we differently process information related to the locations where we look (overt) that from the ones that we do not look (covert) [246]. This distinction can easily be observed by focusing to a particular place in the visual field (e.g. looking at your thumb with your arm stretched forward [177]). Vision away from the central visual field appears to be blurry [303], whereas in the center of view there is full resolution (see Figure 1.2). Given these premises, one could say that the scene context and fixation locations can affect perception of what we are looking at.

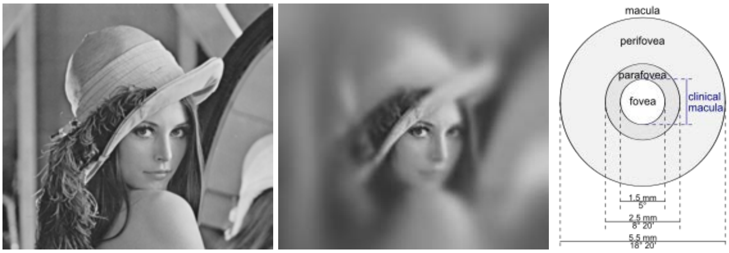


Figure 1.2 – Example of foveal vision. **Left:** Original Image. Middle: Blurred image simulating para/perifoveal vision. **Right:** Illustration of foveal regions in the retina.

In order to decide where to look[1], several human brain mechanisms determine location and time of fixations, as well as amplitude and velocity of saccades [61, 161].

Feature Integration Theory [320] establishes that when searching visual features (e.g. a target with a different orientation, color or size with respect a set of distractors), finding conspicuous objects is efficiently done (in parallel or preattentively). Conversely, in conjunctive search (where a set of distractors have similar combinations of features with respect to the target) it required a serial "binding" step (observing each visual element at a time).

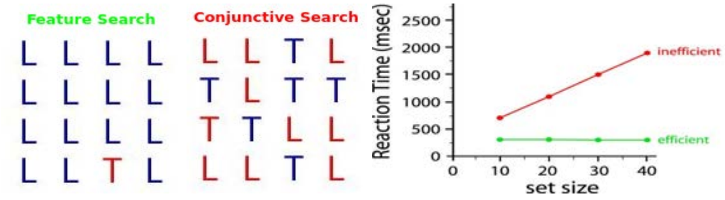


Figure 1.3 – Examples of Feature and Conjunctive Search **(Left-Middle)** stimuli. Finding the Red "T" in Feature Search (green) is easier than for the case of Conjunctive Search (see Reaction Times, **Right**), as we need less fixations to find the target. Adapted from [99]

.

Guided Search model [365] considers that selection mechanisms depend not

[1] Lecture from Zhaoping Li: https://www.youtube.com/watch?v=i2u-5ll5ByA

only on the stimulus itself (bottom-up) but also on decisions of the user (top-down), ranking each feature separately upon stimulus guidance. These two factors (bottom-up and top-down) are thought to compete [84] in order to guide visual attention, selecting those features with higher probability (or priority) [88].

Koch & Ullman [157] came up with the hypothesis that neuronal mechanisms involved in **Selective Visual Attention** generate a unique "master" map from visual scenes, coined with the term "saliency map". In that regard, we tend to move our eyes towards regions that appear to be visually conspicuous or distinct in the scene.

The **Selective Tuning** hypothesis [326] instead suggested that exist multiple saliency maps, which are selected or biased by higher areas in the brain by gating relevant information in each processing level.

The concept of how the human visual system forms a unique priority map (saliency for bottom-up and relevance for top-down) is still an unsolved problem. In summary, eye movements are known to be influenced by saliency, scene context, the task or priming and the internal state of the subject. In this thesis we will discuss how neurons in the cortex are connected and we propose how they could process visual information in order to produce eye movements.

1.1.2 Human Visual System

The human visual system (HVS) is the part of the brain that gives us the capacity to see, in other words, to be able to determine 'what' things are and 'where' they are [110, 333]. Through distinct stages of processing, the HVS is responsible of understanding visual input since the light is projected to the retina, transformed to sensory signals and later processed by the cortex. These steps [108, Chapter 2] can be summarized as:

- **Reception & Transduction**: Retinal Photoreceptors (RP), namely rods and cones, absorb the light that falls onto the retina (corresponding to Long, Medium and Short wavelengths of human visible spectrum).

- **Encoding & Transmission**: Retinal Ganglion Cells (RGC) extract chromatic opponencies from RP signals (red-green, blue-yellow and light-dark) at distinct center-surround polarities (ON/OFF-center) and transmits this information to the Lateral Geniculate Nucleus (LGN) through the optic nerve.

- **Perception & Cognition**: Signals from Parvo-, Konio- and Magno-cellular pathways in LGN are projected to receptive fields in the primary visual cortex (V1 or striate cortex). Neurons in V1 will recurrently process this early visual information and send it to extrastriate ventral (what) and dorsal (where) pathways for higher order processing.

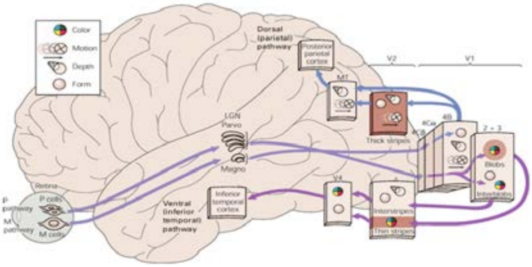


Figure 1.4 – Functional and Anatomical illustration of the HVS from [152, Chapter 25].

Initial neurophysiological experiments [136, 137] discovered that neurons in V1 are sensitive to several properties of the visual stimulus (with electrical single-unit recordings of cat and monkey striate cortex), such as orientation, color, scale, etc. These described our first understanding of how the cortex processes low-level visual features. Cells in V1 receive information from LGN which define their receptive field (RF) activity. It is known that for an ON-center/OFF-surround cell, firing rate is maximal when ON stimulus is located in the center of the neuron RF, and supressed when ON stimulus is located in the surrounding region [348, Chapter 30][200, 299].

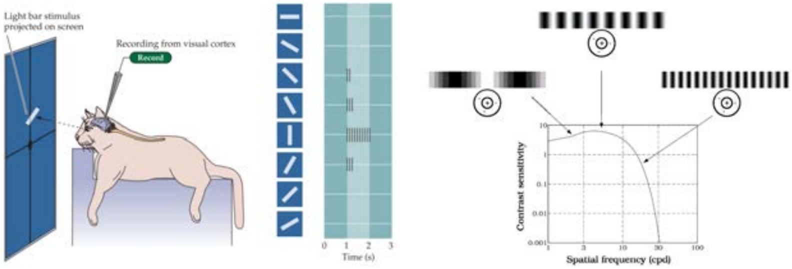


Figure 1.5 – **Left:** Experimental setup from Hubel & Wiesel [136, 137], adapted from [250, Chapter 11]. Neuronal spikes (right panel) from cat's striate cortex appear to be higher for neurons with specific sensitivity to vertical line orientations given the presented stimulus (left panel). **Right:** Contrast Sensitivity Function given sinusoidal grating stimuli [341, Chapter 5]. An ON-center/OFF-surround cell (simulating RGC sensitivities) with constant size responds distinctively to gratings of distinct spatial frequencies.

The HVS encodes retinal information from retinal ganglion cells (RGC) as

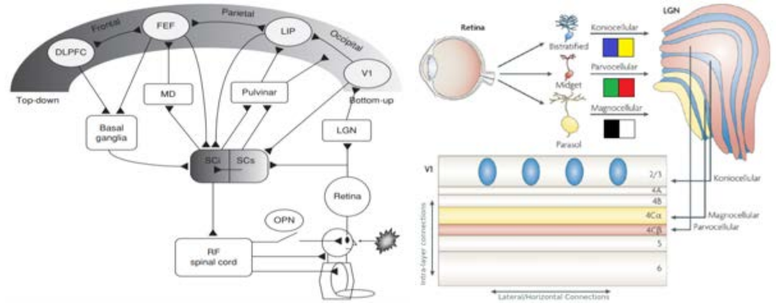


Figure 1.6 – **Left:** Visual areas involved in eye movement control, from [350]. **Right:** Projections from Retina to V1, adapted from [294].

Magno-, Parvo- and Konio-cellular pathways to LGN [Figure 1.6-Right][47, 148, 152, 212, 294]. Cells in LGN transmit ON/OFF activity for each of these chromatic opponencies (light vs dark, red vs green and blue vs yellow) towards specific layers in the primary visual cortex (or striate cortex, V1). Each layer in V1 processes activity from these single cell opponencies in a recurrent manner, given intra-cortical (intra-layer and lateral) and inter-cortical interactions. Feedforward and feedback projections from V1 towards other areas in the brain will determine which percepts and actions will be done upon the scene. More specifically, the superior colliculus (SC) [44, 350] receives activity from distinct cortical areas (V1-V6, LIP, FEF and DLPFC) to trigger voluntary and involuntary eye movements Figure 1.6-Left. Acknowledging the roles of areas that project to SC, locations for fixations depend on the visual representations elucidated by the perceived stimuli, its relation to previously perceived stimuli and its importance with respect to the subject [243, 244, 245, 264].

1.2 Computational Modeling

1.2.1 Visual Hierarchies and Biological inspiration

Some of the challenges in computer vision has been to reproduce perception and tasks performed by humans [124, 156, 164]. Computational models of visual cortex are inspired by simple and complex cell mechanisms [58, 122, 187, 291]. Simple cells are usually modeled as linear filters (either using difference of gaussians or gabor-like filters) to represent retinal center-surround responses [263] and V1 receptive field selectivity to orientations at different spatial scales [37], allowing to represent activity as a pyramid of low-level feature maps. Complex cells are found to combine projections from afferent simple cells, acting as a "pooling" mechanism of the aforementioned low-level features. Receptive fields in higher areas of cortex are bigger and respond to different combinations of features. Arising from that principle, several models of the cortex defined vision as a feedforward mechanism, coined with the term "visual hierarchy".

Fukushima's Neocognitron [100] and Poggio's HMAX [198, 262, 287][Figure 1.7] were the first general hierarchical models, representing the HVS as a chain of pooling mechanisms in order to obtain higher complexity and detail at higher stages of the architecture. HMAX abled to process mid- and high-level visual feature computations, and served for multiple purposes in computer vision (from invariant image recognition [288, 289], shape and texture perception[312, 346] to visual search [334]). However, the HVS is largely known to have both intra-layer (connections between cells in same layer, also named lateral or horizontal connections), inter-layer (connections between cells in different layers), feedforward and feedback connections (Figure 1.6-Right). It makes feedforward-only architectures unable to explain biological and perceptual principles of the brain, making vision yet an unsolved problem [126, 271, 314].

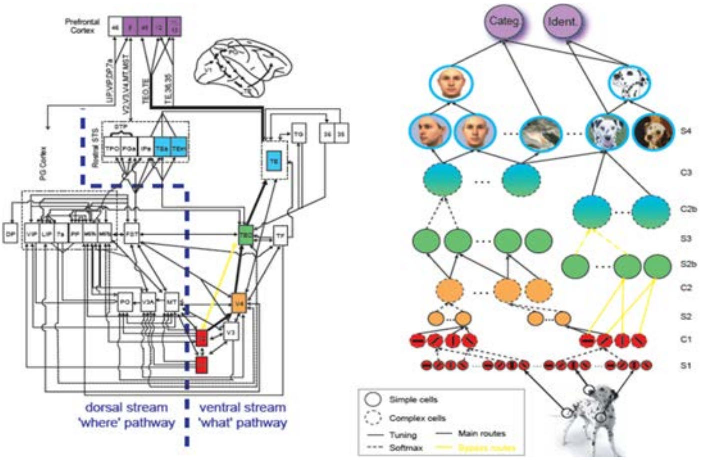


Figure 1.7 – Representation of HMAX computations **(Right)** and its association with function in cortex **(Left)**[288]

Other models such as the Grossberg's LAMINART[114] and Bednar's Cortical Maps [9] proposed to emulate some of these connectivities. Connectivities in LAMINART are constrained by the definition of the architecture, whereas Bednar's model self-organizes from input images. However, the complexity of the visual cortex makes hard to cope implementations of the HVS working with real images and several tasks simultaneously.

Computer vision solved some of the aforementioned problems for real world scenarios, by combining both image processing and machine learning techniques. In line with the efficient coding principle, the statistical relationship between comparing an image patch and a feature basis [191, 315] (e.g. a gabor function or a sparse representation) can define to some extent how objects are quantified by cortical receptive fields [94, 225, 260]. A way to compute this kind of relationship is to obtain a feature map by convolving the image and the kernel basis that best represents each patch of the scene/object that is desired to be computed. However, the key factor here is to define this pattern. One way to do that is to train a model by learning its prediction error with respect the ground truth for each task (supervised) or either uniquely using the data that is available to the model (unsupevised). That principle was used in Artificial Neural Networks (ANN), by processing the input signal in a set of nodes in a feedforward manner and backpropagating errors [273] by weighting each node in the architecture. AlexNet [163] was one of the main architectures that combined these type of computations (convolutions, ANNs and backprop) with real images (also known as Convolutional Neural Networks or CNNs). It outpeformed previous computer vision techniques (as well as human performance in some tasks [104]) for a large variety of fields, and is considered one of the precursors of "Deep Learning" [176, 376][111, Chapter 9] and its latest architectures [6]. CNN architectures are said to work as the brain because some of their computations are to some extent inspired by biological mechanisms [16, 21, 118, 197] (i.e. "convolution" resembling visual feature basis filtering, "pooling" resembling complex cell integration, "rectification" resembling neuronal activation functions and "normalization" by resembling divisive normalization found in cortex). However, we suggest that these type of feedforward architectures cannot solve the vision problem and and simulate the brain [2] because they have:

- No relation of architecture with physiology
- No feedback connections
- No top-down control
- No foveation

Furthermore, they are unable to reproduce:

- Dynamics of visual pathways (alpha, beta and gamma oscillation bands)

[2] Seminar from Simon Thorpe: https://www.youtube.com/watch?v=jKM3S5tMMYo

- Temporal dependency of feature processing
- Psychophysical results

Moreover, the usage of supervision in feedforward architectures require a large number of images to correctly generalize for each concrete task (requiring as well the ground truth for each case), which is an unsolvable problem if we desire to build machines to see and understand the world as a human. In contrast, the HVS does not require that much number of examples for learning objects. As stated in the previous section, humans have ecologically adapted to their environment (considering here that the task of survival contains all the other tasks), knowing "what" things are and "where" things are. This latter problem is constrained by our attention, namely, our eye movements. In the next section we will explain the main computational approaches to predict this phenomena.

1.2.2 State of the art in saliency and scanpath modeling

Visual salience can be defined as “the distinct subjective perceptual quality which makes some items in the world stand out from their neighbors and immediately grab our attention" [143]. In order to understand how visual information is predominantly selected for controlling eye movements, several studies proposed different approaches. Previous theory from Koch & Ullman [157] proposed a computational framework in which visual features are integrated to generate a saliency map. These visual features are projected to V1 and later processed distinctively on the ventral (“what") and dorsal (“where") streams. These connections are projected to the superior colliculus (SC), which would generate either top-down (relevance) or bottom-up (saliency) control of eye movements by combining neuronal activity from distinct brain areas to a unique map (priority map) [88][350]. Given these distinct levels of processing, a set of computational models are proposed in order to reproduce eye movement behavior. Itti et al. introduce a biologically-inspired model [142]Figure 1.8-Left in which low-level features are extracted using linear DoG filters, their conspicuity is calculated using center-surround differences (inspired by V1’s simple cell computations) and integrated (pooled to the SC as a master saliency map) using winner-take-all (WTA) mechanisms.

Although computations of existing saliency models seem to mimic HVS mechanisms, complexity of scenes make eye-movement behavior hard to predict. Bruce & Tsotsos model [50][Figure 1.8-Right] offered a semi-supervised mechanism to account for relevant information of the scenes in combination with the bottom-up computations of V1, predicting eye movement behavior at distinct scene contexts. Given the basis of these models, a myriad of computational models, both with artificial and biological inspiration [150][41][379][257], have implemented distinct ways

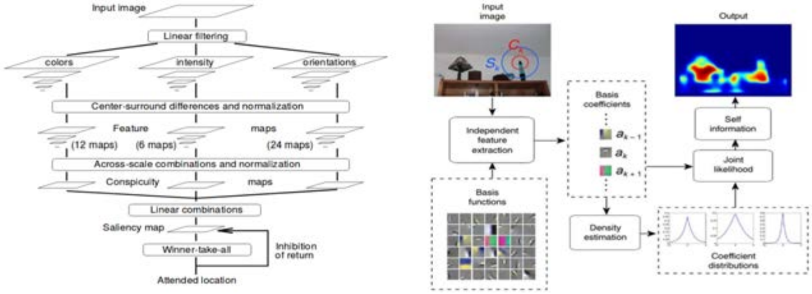


Figure 1.8 – **Left:** IKN [142] extracts feature maps with DoG, then computes center-surround differences and integrates them to the saliency map with WTA. **Right:** AIM [50] instead convolves the image with kernels that maximize visual information (through ICA of patches from a set of 3600 natural images) and estimates the joint likelihood to obtain the saliency map.

to predict human eye movements obtaining better performance on its predictions [256][43][46][53]. By processing global and/or local image fatures for calculating feature conspicuity, these models are able to generate a master saliency map to predict human fixations (Table 1.1). Taking up Judd et al. [150] and Borji & Itti's [41] reviews, 5 general categories of model inspiration follow similar saliency computations:

(C) **Cognitive/Biological:** Saliency is usually generated by mimicking HVS neuronal mechanisms or either specific patterns found in human eye movement behavior. Feature extraction is generally based on Gabor-like filters and its integration with WTA-like mechanisms.

(I) **Information-Theoretic:** These models compute saliency by selecting the regions that maximize visual information of scenes.

(P) **Probabilistic:** Probabilistic models generate saliency by optimizing the probability of performing certain tasks and/or finding certain patterns. These models use graphs, bayesian, decision-theoretic and other approaches for their computations.

(F) **Spectral/Fourier-based:** Spectral Analysis or Fourier-based models derive saliency by extracting or manipulating features in the frequency domain (e.g. spectral frequency or phase).

(D) **Machine/Deep Learning:** These techniques are based on training existing machine/deep learning architectures (e.g. CNN, RNN, GAN...) by minimizing

Table 1.1 – Description of saliency models.

Model	Authors	Year	Inspiration					Type	
			C	I	P	F	D	G	L
IKN	Itti et al.[142, 145]	1998	✓					✓	✓
AIM	Bruce & Tsotsos [50]	2005	✓	✓					✓
GBVS	Harel et al.[119]	2006			✓			✓	✓
SDLF	Torralba et al. [316]	2006			✓			✓	✓
SR & PFT	Hou & Zhang[131]	2007				✓		✓	
PQFT	Guo & Zhang[116]	2008				✓		✓	
ICL	Hou & Zhang [132]	2008		✓	✓			✓	✓
SUN	Zhang et al. [380]	2008			✓				✓
SDSR	Seo & Milanfar [286]	2009	✓		✓			✓	✓
FT	Achanta et al.[2]	2009				✓		✓	
DCTS/SIGS	Hou et al.[130]	2011				✓		✓	
SIM	Murray et al.[209]	2011	✓					✓	✓
WMAP	Lopez-Garcia et al.[188]	2011	✓			✓		✓	✓
AWS	Garcia-Diaz et al.[102]	2012	✓					✓	✓
CASD	Goferman et al.[107]	2012	✓	✓	✓		✓	✓	✓
RARE	Riche et al.[259]	2012		✓				✓	✓
QDCT	Schauerte et al.[279]	2012				✓		✓	
HFT	Li et al.[182]	2013				✓		✓	
BMS	Zhang & Sclaroff [377]	2013			✓			✓	
SALICON	Jiang et al.[147, 313]	2015					✓		✓
ML-Net	Cornia et al.[75]	2016					✓		✓
DeepGazeII	Kümmerer et al.[168]	2016					✓		✓
SalGAN	Pan et al.[230]	2017					✓		✓
ICF	Kümmerer et al.[168]	2017			✓			✓	✓
SAM	Cornia et al.[77]	2018					✓		✓
NSWAM	Berga & Otazu [26]	2018	✓					✓	✓

Inspiration: { C : Cognitive/Biological, I : Information-Theoretic, P : Probabilistic, F : Fourier/Spectral, D : Machine/Deep Learning} Type: {G: Global, L: Local}

the error of predicting fixations of images from existing eye tracking data or labeled salient regions.

Whether proposed computational saliency models resemble eye-tracking data, it is questionable to consider that these predictions accurately and specifically represent saliency [51][23]. While the current concept of saliency maps is to predict probabilities of specific spatial locations as candidates of eye movements, it is also crucial to understand how to predict individual fixations or saccade sequences (also named "scanpaths"). Scanpath predictions can be done through probabilistic measures of saccade amplitude statistics. These followed a similar heavy-tailed distribution similar to a Cauchy-Levy one (in reference to random walks or "Levy flights", minimizing global uncertainty) [48], with highest probability of fixations at a low saccade amplitude. This procedure was implemented in Boccignone & Ferraro's scanpath model[39], using saliency from IKN. Later, LeMeur & Liu[202][Figure 1.9] proposed a more biologically plausible approach, accounting for oculomotor biases and inhibition of return effects. Latest scanpath model (STAR-FC) from

Wloka et al. [357][Figure 1.10] included an excentricity-dependent foveation mechanism reproducing retinal acuity [345], then cropped the fovea center in order to process low-level saliency and high-level saliency for central and peripheral maps respectively.

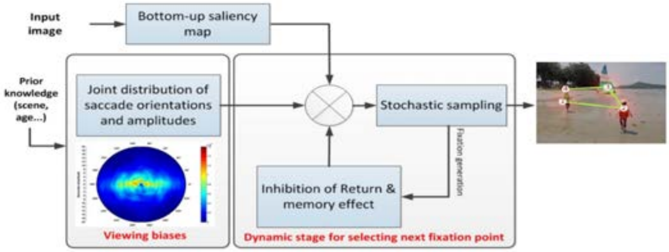

Input image
Bottom-up saliency map
Prior knowledge (scene, age...)
Joint distribution of saccade orientations and amplitudes
Stochastic sampling
Fixation generation
Inhibition of Return & memory effect
Viewing biases
Dynamic stage for selecting next fixation point

Figure 1.9 – Pipeline from LeMeur et al.'s scanpath prediction model [202], from [201]. By combining the saliency map (from GBVS [119]) with Saccade Amplitude and Orientation statistics, as well as a simulation of Inhibition of Return is able to predict saccade sequences.

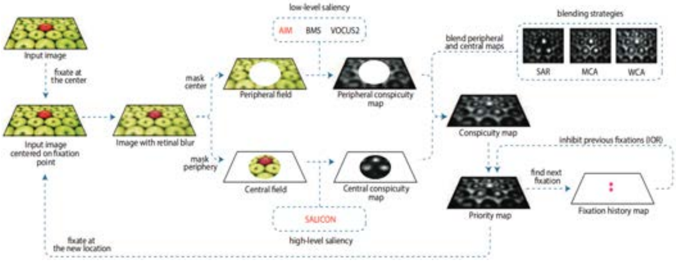


Figure 1.10 – Pipeline from Wloka et al.'s Selective Tuning Attentive Reference Fixation Controller model, from [357]. It follows previous architecture by Tsotsos et al. [327], with inhibition of return mechanisms and distinct saliency computations for central and peripheral fields, joined to the priority map as targets for fixations.

Psychophysics of visual attention

Part I

Measuring eye movement behavior.

2 Psychophysical evaluation of individual low-level feature influences in visual attention

Visual attention is the cognitive capacity of efficiently selecting relevant visual information from a scene. Researchers record eye movements in psychophysical experiments using eye-tracking technology as means of identifying overt attentional cues around fixation points, [161]. Registered data of different subjects show different patterns of eye movement depending on reflexive, goal-directed or contextually-specific influences [270]. This suggests the existence of two types of general influences in the Human Visual System (HVS), combining both bottom-up and top-down processing [84][172][73][93][350]. Bottom-up processing of low-level visual features takes place in the early stages of the HVS, namely, when the nervous system efficiently extracts the basic information of the scene and processes it in the visual cortex. When higher areas of the brain are involved is when the top-down processing occurs, by taking into account internal state of the subject (task, mental state, experiences, etc.).

2.1 Problem Statement

The limits on the prediction capability of saliency models [Chapter 1.2.2] arise as a consequence of the evaluation from previous datasets, that do not account contextual, perceptual, temporal and task-related biases.

2.1.1 Contextual Relevance

One of the properties that guide visual attention is the contextual relevance of the observed scene [229][68][123][235][339][140][81]. Semantically-relevant content or specific high-level features can generate endogenous attentional guidance. For instance, looking at a website promotes specific eye movement patterns that differ from looking at a nature scene image; different scanpath patterns can also be found in eye-tracking experiments while humans observe indoor, outdoor and synthetic images. In most datasets for saliency modeling, observers perform free-viewing tasks with real images labeled in specific scene context categories (either faces, cars...), without taking full account of the top-down priors influenced by the context of the image with respect to feature contrast [355][43], which could bias both feature localization and discrimination difficulty [228][337][366].

2.1.2 Contrast Relevance

Eye movement behavior is influenced not only by content and stimulus context, but also by the human perceptual capabilities for distinct contrast adaptation and discrimination [218][238][242][196][120]. Other perceptually-relevant factors could also be related either to the lighting conditions used in each experiment, the starting point of view when perceiving stimuli, etc. The evaluation of relative distinctiveness between features at distinct regions of the image is needed to be done in order to analyze each image according to its spatial properties and feature specificities. This suggests that each image promotes distinct saliency.

2.1.3 Temporal Relevance

Eye movements have been shown to have temporal influences, varying its behavior upon viewing time or number of fixations (e.g. showing decreasing saccade amplitude, increasing fixation duration [98][8] or higher inter-participant differences [306][269]), suggesting the idea that saliency influences more early saccades than late viewing saccades [234][306][381][386]. Most saliency predictions based on eye tracking data do not evaluate the temporal relevance in relation to the saliency elicited by the scene, being for most cases, evaluated spatially across all fixations.

2.1.4 Task Relevance

Alfred Yarbus' seminal work revealed differences in eye movement patterns [373] caused by certain top-down influences such as previous experience, motivation and other endogenous factors. Distinctive studies have also concluded that task priors are decisive in that respect [52][213][307][64] [113][40]. Goal-directed tasks proved to be able to condition eye movement behavior enhancing visual attention processing [246][149][138][170]. That might suggest that visual search tasks could minimize such eye-movement patterns produced by endogenous top-down mechanisms [129][360], by increasing induced attention towards salient targets (combining both saliency and relevance to influence eye guidance towards these regions). Thus, for all tasks, there is an induced top-down processing that tune overall visual priority when recording eye-movements [123] [149][83], given both exogenous and endogenous influences. Such design puts forward that there could be a better computational estimation of saliency if such task-related influences were focused uniquely on the regions that pop-out on the scene.

2.1.5 Center bias

Eye movement datasets built for the assessment of saliency models tend to be center biased, not only because of scene framing (photographies tend to focus the salient region in the center of view) but also because of the specific task and stimuli, whereof top-down modulatory constraints are enough to prevent attentional shift, giving a trend to promote center biases [45][192][338][71][269], not only in oculomotor terms but also in tendencies in experimentation of eye movement behavior. As aforementioned, bottom-up and top-down processing of the stimuli will depend on the feature characteristics from the scene. If these are simpler, the contextual influence will be lower, making the indicators of saliency easier to analyze [305][338]. There will be an endogenous top-down attentional modulation whether the stimulus is cued or uncued. For concrete salient stimuli, facilitating attentional guidance by inducing specific endogenous cues could enable the selection of specific regions of interest in order to prevent the aforementioned factors that generate these center biases.

2.2 Objectives

Acknowledging the aforementioned problems on capturing bottom-up visual saliency, we have decided to create a dataset with synthetic images, lacking the presence of high-level features, promoting saliency uniquely elicited from low-level features (providing as well a synthetic image generator code). An alternative evaluation of saliency proposed, by measuring eye movements upon low-level feature distinctiveness and their temporality. Fixations and saccades will be evaluated individually with the corresponding stimuli on free-viewing and visual search tasks, with different feature types and distinct target-distractor feature contrasts.

In order to vary the level of saliency of specific features in a scene, a parametrization of the distinctiveness between a specific item and a set of distractors or its surrounding background is needed. By parameterizing feature contrast, it is possible to analyze feature search efficiency, its accordance with the Weber Law, and the effects in which search asymmetries apply. Using synthetic images in eye-tracking experiments, the complexity of the image features is reduced by minimizing any top-down contextually-related effect, putting forward an easier and more accurate evaluation of eye movement behavior. By modeling stimulus areas of interest for selected pop-out targets, it it possible to test participants performance on landing inside salient regions and their eye movement patterns (in the extent of fixation duration and saccade amplitude) for distinct feature contrasts, and their temporal evolution. This will allow us to observe whether low-level features influence visual attention in a distinct manner.

Previous experiments that perform psychophysical tests (see Table 3.1) evaluating human visual performance on distinct low-level features (iLab USC [145], UCL [385], VAL Harvard [367] and ADA KCL [300]) show that the distinctiveness between a specific region and the rest of distinct regions of an image progressively increases the level of saliency in relation to feature contrast. However, the presence of much less relevant features distorts the overall distinctiveness of a specific region with respect to the rest, thus, affecting to the bottom-up visual guidance towards the salient region. With the aforementioned datasets, feature contrast and stimulus conditions has been parametrized with search tasks (using the button trigger for calculating search reaction times) but no eye tracking experimentation has been done.

For few eye movement datasets that contain synthetic images (MIT[151] and CAT2000 [42]), no parametrization of feature type or contrast was done. Contrary to other saliency datasets [54], in this study it is possible to evaluate each of these factors individually and exclusively eye movement data is being used for calculating search performance for better accuracy. We will test the following hypotheses:

1. Performance on salient region localization could show differences upon varying the type of features present in our stimuli.

2. If feature contrast is the main factor that contributes to saliency, performance on localization of salient regions should correlate with feature contrast, specially for stimuli that require a serial 'binding' step.

3. Acknowledging that saliency is usually evaluated across all fixations on eye tracking experimentation, if a temporal bias exists and is increasing, it is highly possible that the first fixations show higher saliency index than the late ones.

4. If performance on salient region localization with free viewing tasks is lower for stimuli with higher contrast compared to visual search tasks, it will mean that fixations on free viewing tasks are highly guided by endogenous attention.

5. Previous datasets used for saliency prediction do not show how their center biases affects saliency. We will show how eye movement patterns influence the center bias for this dataset and if the bias increases or decreases across viewing time and feature contrast.

Our objective is to allow computer vision researchers to reproduce these influences when modeling eye-movement prediction algorithms. Here we present a dataset in which we evaluate through free-viewing tasks the influence of the features that affect the spatial properties of an image (from the perception of Corners, Segments, Contours and Grouping) and how the relative distinctiveness from a search target is more salient with respect to a set of distractors that differ from specific low-level features (color, orientation, size...). Analyzing low-level features individually would allow us to see which features generate more agreement on saliency measures and are localized faster, in that manner, to allow their modeling according to their distinct neuronal mechanisms. This study can be used for a more plausible and specific saliency modeling given the presented eye-movement patterns, also extrapolable to the analysis of the interactions between these features or to study specific cases of high-level features in future studies.

2.3 Stimuli

A total of 33 types of stimuli were generated, corresponding to 15 distinct feature evaluations (5 of them using free-viewing tasks and 10 for visual search tasks) at distinct conditions. During free-viewing experiments, we evaluated how spatial properties influence saliency, namely, the capabilities of humans for detecting corners, segmenting and detecting contours as well as localizing groups of objects according to their similarity and spatial distribution (Table 2.1). This will give some insight of how rapidly humans reflexively perceive and bind spatial properties from the features of an image. In visual search tasks, we evaluated the speed in detecting specific features and the amount of saliency produced by target-distractor feature contrast characteristics. In that aspect, stimulus were generated with features that pop-out based on their dissimilarities in orientation, color and size. Besides, we also analyzed influences of the guidance prompt from the amount of distractors on the scene, their configuration as well as the influence of background lightness, color and roughness (Table 2.2).

Stimulus design was was inspired by Spratling's experiments [300], by generating synthetic images similar to the ones from Li and May's psychophysical experiments [385]. Most stimuli items had a size of 1.5 deg, occupying a region of 2.5 deg including the spacing between distractors. In that manner, stimulus had an available grid of 10 × 13 distractors. Distractors were black ($lsY = 0,0,0$), and background was plane white ($lsY = 0.6548, 0.0175, 1$). We used Spratling's code and we adapted it in order to also use any distractor shape, displacement and chromatic parameterization.

$$\Psi(x) = \{\frac{x-1}{N-1} \quad | \quad x \in N\}, \tag{2.1}$$

$$\Psi(x, v) = \{v \cdot \Psi(x)\}, \tag{2.2}$$

$$\Psi(x, min, max) = \{\Psi(x, min) \cup \Psi(x, max) \quad |\text{x is odd}\}. \tag{2.3}$$

The parameters of the generated stimuli were set according to "$N = 7$" contrast values (ranging from 0 to 1), using the Weber's law uniform fraction in order to set discrete target-distractor evaluation "$x = 1...N$" for the psychometric function Ψ. For each stimulus type on our experimentation, parameters are set according to specific values of $\Psi(x)$. We have the expression in Equation 2.1 and 7 contrast values in order to have extreme contrast values (no contrast and maximum contrast) with "$\Psi(1) = 0$" and "$\Psi(7) = 1$" as well as a middle value with "$\Psi(4) = 0.5$", making the difference between the second lower contrast and the second maximum contrast at the same distance from the extreme contrast values $|\Psi(2) - \Psi(1)| = |\Psi(7) - \Psi(6)|$. In that manner we provide a psychometric function with a constant slope (Weber's uniform fraction). Absolute values of contrast can be adjusted to fit a specific value of "v" as $\Psi(x, v)$ (Equation 2.2). For cases that we had higher and lower contrasts with respect the target and overall distractors we adjusted the values for maximum and minimum range of the psychometric function $\Psi(x, min, max)$ as the union of odd values for both sets of $\Psi(x, min)$ and $\Psi(x, max)$ in order to acquire the same set of contrast values (Equation 2.3).

Acknowledging that each stimulus was distributed according to different contrasts depending on the evaluation parameter, each stimulus was categorized as easy and hard depending on the assigned contrast (half of them as easy for higher contrasts, and half of them as hard for the case of lower contrasts, with a specific case with minimum or no contrast). One of our interests was to evaluate how low-level features modify the spatial layout between the features on a scene, therefore its spatial properties, affecting visual saliency (in this case with free-viewing experimentation). In order to accurately evaluate low-level feature distinctiveness, visual search tasks were performed, having a search target with a specific low-level contrast with respect to a set of distractors. The stimulus design corresponding distinctively to each feature and task will be explained as follows.

Free-viewing task stimuli

First, we wanted to evaluate the spatial relevance of certain regions of an image. For this stimuli, visual selection cannot be focused on a unique region due to the size and/or spatial organization of the elements in the image. Humans have a limited central vision, namely, they need several fixations over the whole region in order to attend to all of the relevant regions in detail. In that aspect, each of the spatial properties will guide attention towards a single or several spots depending on the analyzed feature. With this type of stimuli it is be able to see the temporal and spatial performance of perceiving boundaries due to corner sharpness, segment angle and spacing as well as preemption and grouping [254, 255, 368], effects induced by distractor continuity, proximity and similarity. These preattentive effects are not equally processed in the visual system in the same way as shown for parallel visual search [361]. Task was separate for the aforementioned perceptual phenomena with respect to searching for a specific feature, stimuli described on Section 2.3.

Table 2.1 – Description of the generated stimuli for the experiment using the free-viewing task. Stimulus have been divided in "Stimulus type" according to the type of feature or effect that is analyzed and "Stimulus subtypes" for the cases that there are presented distinct conditions using the same feature contrast. The total number of elements has been selected according to the stimulus characteristics, preserving similar spatial properties to the ones presented on the literature.

# of stimuli	Stimulus type	Stimulus subtypes	Parametrized Feature Contrast	Total # of elements
7	Corner Angle (1)		Sharpness Orientation	1
14	Segmentation by Angle (2)	Single Superimposed	Segment Orientation	10 × 13 (130) 20 × 26 (520)
7	Segmentation by Spacing (3)		Bar Length and Spacing	10 × 13 (130)
6	Contour Integration (4)		Bar Continuity	10 × 13 (130)
14	Perceptual Grouping (5)	Similar Dissimilar	Distractor Proximity	~40

Corner Angle (1) Troncoso et al's psychophysical experimentation found that corner salience was higher on sharp corners than on shallow corners or edges [323, 324]. This effect could be explained by ON-center receptive field behavior towards corner stimuli [263], being sharp corners the ones that produce higher neuronal activity. Original stimuli from Troncoso's experiment was used, generating corners with a dark-to-white gradient and an upwards angle, corresponding to corner angles of $180, 135, 105, 75, 45, 30$ and 15^{o} (shown in Figure 2.1). The horizontal alignment of the corner was randomized in order to prevent oculomotor anticipation.

Visual Segmentation Distinctiveness between two homogeneous regions creates higher neural activity near region boundaries than away from them [185, 186]. In this section is described how an illusory boundary is generated by varying two segment characteristics. This effect is distinct from the concepts of edge or boundary detection (terms used as well in the image segmentation literature) or contour

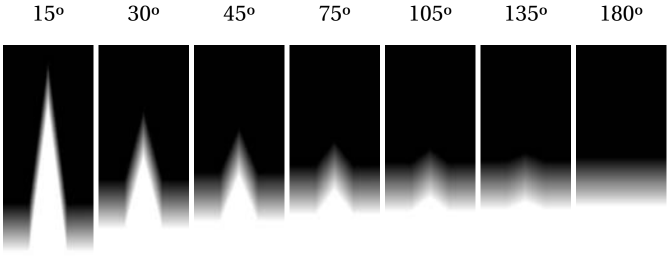


Figure 2.1 – Examples of corner angle slopes (with the sharper at 15 º and the smoother at 180 º) for dark-to-bright gradient stimuli with upwards angles.

integration. This phenomena proves that illusory boundaries pop-out due to the perceptual breakdown of homogeneity. Here it is studied the influence of angle contrast between these two segments (creating a salient boundary dependent on the segments angle) with an homogeneous single set of bars as well as with superimposed bars. Here is also analyzed the influence of bar spacing and length on detecting the illusory boundary between these two segments.

Segmentation by Angle (2) Visual angle contrast between two segments can induce edge detection and therefore visual saliency towards that illusory edge [222, 369, 385]. The resulting saliency would increase with respect to angle contrast from the two segments on the region that separates them [300]. It is a distinctive effect from orientation feature detection upon a set of distractors, that is described on Orientation Contrast (12), Distractor Heterogeneity (13), Distractor Linearity (14) and Distractor Categorization (15).

$$\Phi(v,a) = \{|\arcsin(\Psi(1...N,v)) + a|\}, \tag{2.4}$$

$$\Delta\Phi(v,a,b) = min\{|b-\Phi(v,a)| \quad , \quad 180-|b-\Phi(v,a)|\}, \tag{2.5}$$

The psychometric values for determining angle values are defined as $\Phi(v,a)$. Here "v" is the incremental factor for adjusting the maximum angle for our set $\Psi(x,v)$ and "a" is the starting angle value for our bar orientation (Equation 2.4). The angle contrast between a specific angle "b" and our set of angles $\Phi(v,a)$ can be computed with $\Delta\Phi(v,a,b)$, considering that our bar orientations have upwards and downwards contrast for its comparison (due to its symmetry), contrast is calculated as the minimum from the differences from two quadrants in which these bars can be oriented (Equation 2.5).

Stimuli was based on Spratling's visual segmentation, using 2 sets of bars (shown in Figure 2.2(**a,b**)) oriented respectively using angles $\Phi(1,0)$ and "$b = 90$", forming a relative contrast of $\Delta\Phi(1,0,90)$. For the case of superimposed bars, we have created a composite of the same bars adding a bar tilted at 45 º with respect to each segment. Here are accounted the contrasts between the new superimposed bars and the original segment $\Delta\Phi(1,45,90)$. The location of the vertical segment was randomized on the horizontal axis for each stimulus.

Segmentation by Distance (3) Texture discrimination was shown to vary according to the spacing and length of the texture elements [221, 300], making it harder as element spacing increases (as when segment elements decrease in length or size). Visual segments were modeled using 2 sets of bars, oriented respectively at 45 º and −45 º (a relative angle contrast of 90 º). Here we question how bar length is able to generate a specified distance at the center of the illusory segment. Segment spacing was calculated as the euclidean distance from the end of the first segment bar to the beginning of the second segment bar, with values of 0 to 2.5 deg (shown in Figure 2.2(**c**)), corresponding respectively to a bar length of 1 to 3.6 deg deg in the horizontal axis.

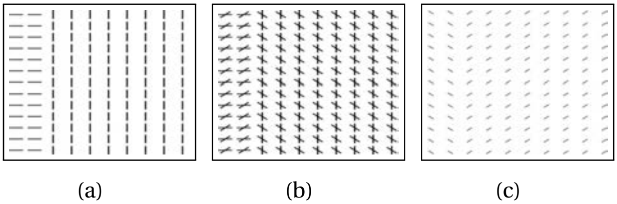

(a) (b) (c)

Figure 2.2 – Examples for visual segmentation stimuli. **(a)** Corresponds to the segmentation by angle with a single segment and **(b)** to superimposed segments (with both cases with an orientation contrast of ΔΦ=90 º. In **(c)** segmentation is done distinctively (by changing bar length), using bars oriented at 45 and −45 º with a bar length of 1 deg and a segment spacing of 2.5 deg.

Perceptual organization Perceptual organization has been previously investigated and promoted by Gestalt principles, guided by proximity, similarity, continuity, and closure properties of objects [57, 94, 158, 272, 349]. Here are described two effects related to perceptual organization, parametrized upon the aforementioned principles.

Contour Integration (4) Continuity within set of features in a scene is able to generate the perception of a contour [125], considering that a larger set of collinear bars facilitate its detection. Accounting for saliency being influenced by contour integration [78, 183, 300], a set of stimuli was created with a grid of randomly oriented and equidistant bars and a collinear contour (Figure 2.3(**a**)). Contours were generated with a length of $3,5,7,8,9$ and 10 collinear bars, corresponding to 7.5 to 25 deg.

Perceptual Grouping (5) Adding up to the basis of the previous section, we have also studied the relation between perceptual grouping principles and visual attention. According to the literature, the spatial layout can facilitate or prevent contextual cueing [34, 72], in particular, the lower the proximity between a number of randomly distributed objects and a group, the higher the saliency on the grouping region [19, 20, 221, 261]. Given that, here the analysis is on the influence of proximity and similarity among objects in a specific spatial organization. To do so, there were generated a set of shapes, randomly distributed and located at specific distances to a group, with similar and dissimilar shapes Figure 2.3(**b,c**). The proximity parameter was the euclidean distance between the group centroid and the rest of distractors, forming a wider gap between the distractors and the group as we increase distance, parametrized as $\Psi(1...N, 2.5, 7.5)$. Stimulus shapes were set to be symmetric in order to prevent orientation-variant guidance, with squares as the main shape for both group and distractors in the case of similar object condition, whereas in the case of dissimilar condition were selected triangle shapes for the distractors and squares for the group.

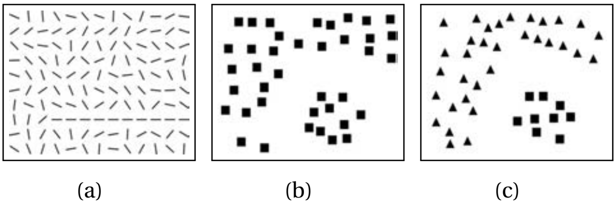


Figure 2.3 – Examples of distinct perceptual organization effects, eliciting Contour Integration **(a)** formed by 10 collinear bars (corresponding to 25 deg) and Perceptual Grouping for similar **(b)** and dissimilar shapes **(c)** with respect a group set at a distance of 7.5 deg from the rest of distractors

Visual Search task stimuli

Visual search tasks were performed with another set of stimuli. In this case, a unique target item with a specific size was used. In that manner, it is possible to change either the amount of distractors, their spatial configuration, the target-distractor feature contrast and background properties of the stimulus. Most target elements overall occupied a small area of interest in order to able to be preserve same fovea-dependent capabilities for each type of stimuli used with this type of task. Hence, using this experimentation we can observe which features pop-out faster and more often (in parallel or "effortlessly"). As small regions away from central vision cannot be detected as in the case of bigger regions shown on Section 2.3, the guidance towards the salient target (distinctive from the rest of distractors) minimizes other type of guidance promoted from endogenous factors.

Table 2.2 – Description of the generated stimuli for the experiment using the Visual Search task. The total number of elements has been selected according to the amount of distractors presented on the scene acknowledging that one of the elements is presented to be the search target.

# of stimuli	Stimulus type	Stimulus subtypes	Parametrized Feature Contrast	Total # of elements
28	Feature Search (6)	Feature Conjunctive Feature-absent Conjunctive-absent	Distractor number	3 to 35
14	Search Asymmetries (7)	Bar presence Bar absence	Scale and Distractor number	35 to 520
14	Noise/Roughness (8)	Higher deviation Lower deviation	Surface Roughness	1
28	Color Contrast (9)	Red target and Unsaturated Background Red target and Oversaturated Background Red target and Unsaturated Background Blue target and Oversaturated Background	Distractor Saturation	34
14	Brightness Contrast (10)	Light Background Dark Background	Distractor lightness	34
7	Size Contrast (11)		Target Size	34
7	Orientation Contrast (12)		Target Orientation	34
21	Distractor Heterogeneity (13)	Homogeneous Tilted-right Flanking	Target Orientation	10 × 13 (130)
28	Distractor Linearity (14)	Linear Nonlinear at 10º of slope Nonlinear at 20º of slope Nonlinear at 90º of slope	Target Orientation	10 × 13 (130)
21	Distractor Categorization (15)	Steep Steepest Steep-right	Target Orientation	10 × 13 (130)

Feature and Conjunctive Search (6) Feature search increases probability and efficiency of saccading towards a specific search target on scene observation due to its unique distinctiveness. The information span processed by the HVS varies depending on the amount of feature distractors to be processed [121, 218, 227, 237, 321, 364, 365]. Given that premise, the amount of objects in a scene would imply a variation of the difficulty towards searching a specific target for the case of serial search (distorting human's sustained attention), but not for the case of parallel search. Previous experiments show that difficulty on visual search is higher with a conjunction of distractors with different image features (such as size, color or orientation). In case distractors vary only by a unique feature, the difficulty of the task would not be as evident as the other case [97, 211, 317, 362]. In order to reproduce feature and conjunctive search, target was a red bar oriented at 45 º. For the feature search case, distractors were green and set at 45 º (Figure 2.4). On the case of conjunctive search, half of distractors were green and set at 45 º and the other half were red and oriented at −45 º.

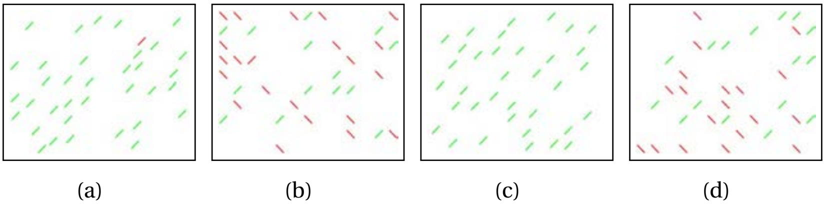


Figure 2.4 – Examples used for feature and conjunctive search. Here are presented the cases of having a red target oriented at 45 º and 34 distractors randomly displaced around the scene. For the feature search case **(a)**, all of the distractors are distinct in color (green). For the conjunctive search case **(b)**, 50 % of the distractors are distinct in color (green) and the rest are distinctive in orientation (at an orientation of −45 º). The same cases **(c)** and **(d)** are shown but without the presence of the target.

The position of the items was randomised, with a set size of $\Psi(1...N,2,34)$, the amount of distractors ranged from 2 to 34 distractors. Both search conditions without the presence of the target was introduced in order to see if the effects are also reproduced for the case of reporting absence of target. The design of feature and conjunction search has been defined as keeping similar difficulty between the two conditions, preserving identical targets and displaying each conjunction of distractors maximally dissimilar from each other [251].

Search Asymmetries (7) Search asymmetries between two different type of stimuli happen when a specific target of type "a" is found efficiently among distractors of type "b", but not in the opposite case (searching for "b" among distractors of type "a") [121, 319, 320, 363]. Clear evidence was found for plain circles crossed by a vertical bar (Figure 2.5) at a scale of 5 deg), showing that it was easier to find a circle with a vertical bar among plain circles than vice versa [300, 363].

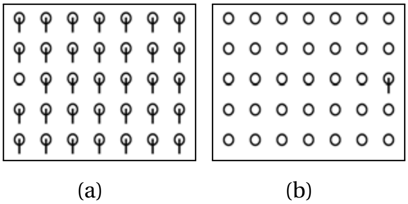


Figure 2.5 – Example of stimulus types in which search asymmetries can apply. The context of circles **(b)** facilitates the search of a superimposed bar compared to the reverse case **(a)**.

The same type of stimuli was selected (with both conditions: searching a circle crossed by vertical bar among plain circles and searching a plain circle among circles crossed by a vertical bar) filling a grid of distractors according to a specific scale and randomizing the position of the target. The scale values were $\Psi(1...N, 1.25, 5)$, between 1.25 and 5 deg, changing the amount of items to be presented, being in each case from 35 to 520 elements corresponding to arrays of 5 × 7, 6 × 8, 8 × 10, 10 × 13, 15 × 20 and 20 × 26 objects.

Noise/Roughness (8) For most synthetic stimuli we have considered uniform and plain backgrounds with homogeneous illumination, but in this case the influence of continuous textured background would increase or reduce search time required to detect a specific target depending on the amount of noise present in the scene. Clarke et. al. [69, 70, 226] showed that the higher the level of background texture noise of a scene, the higher the level of difficulty of the search task. They represented the background surface as a height map by parameterizing an isotropic and random-phase noise $1/f^{\beta}$ (being "β" the frequency roll-off magnitude factor of the inverse discrete Fourier transform of the height map and "σ_{RMS}" the deviation of the roughness noise height). The surface was obtained by rendering the height map according to the Lambert's Cosine Law model using a constant light source with slant of 60 º and tilt equal to 90 º. Given these previous experiments, each stimulus was a rough surface considering β as the contrast value $\Psi(1...N, 1.5, 1.8)$ with two distinct conditions by using deviations of σ_{RMS}= 0.9 and 1.1. A similar target of Clarke's experimentation was used (Figure 2.6) with a circular shape and a vertical gradient background corresponding to the height of the surface and a diameter of 0.78 deg (half of size corresponding to the rest of target items of this study, adjusted for preventing too low RT differences between distinct contrasts).

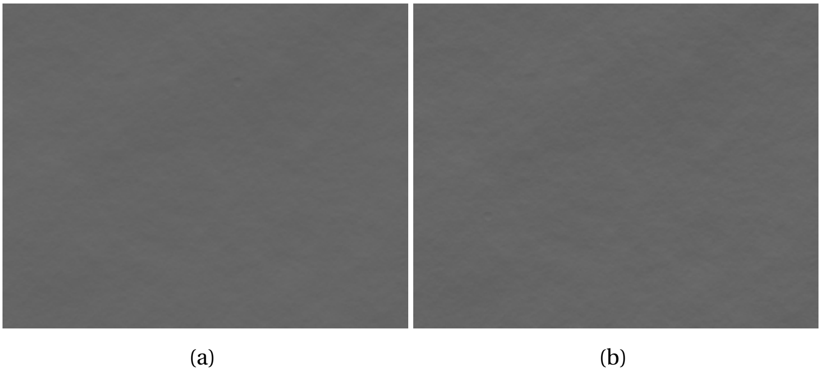


Figure 2.6 – Two examples of a rough surface with $\beta = 1.8$ using height deviations of **(a)** $\sigma_{RMS} = 0.9$ and **(b)** $\sigma_{RMS} = 1.1$

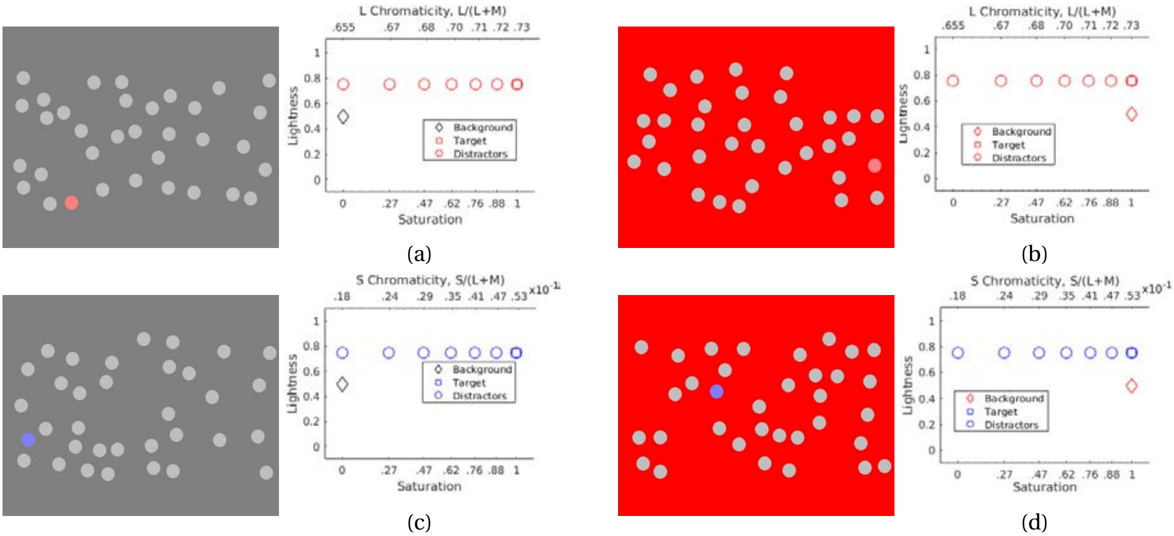
Distractor similarity When an object is dissimilar to the rest of objects in a scene, the search of that object is more efficient. That phenomenon is called target-distractor similarity, and has been found to occur when parameterizing specific features such as color, shape or size [86, 359].

Color Contrast (9) In this section the chromatic properties of distractors are changed, as well as the background of the stimuli Figure 2.7. As shown in previous experiments [7, 17, 79, 87], color varies spatial and temporal patterns of eye movements, affecting both localization and discrimination of objects. Besides, search asymmetries happen at different background conditions [210, 268].

$$\begin{aligned} \Delta S_{1,2} &= |S_1 - S_2|, \\ \Delta L_{1,2} &= |L_1 - L_2|, \end{aligned} \tag{2.6}$$

$$\alpha = \arctan(\frac{\Delta L_{T,B}}{\Delta S_{T,B}}), \tag{2.7}$$

$$\theta = (90 - \alpha)(\Psi(1...N)), \tag{2.8}$$



(a) (b) (c) (d)

Figure 2.7 – Examples of the 4 conditions at maximum contrast of $\Delta S_{D,T}$=1, representing the values of Saturation and Lightness on each particular stimuli for each target and background configuration (showing as well the "l" and "s" chromaticities in the lsY space [189] at 400 nm). In **(a)** and **(b)** there are represented the stimulus for grey (unsaturated) and red (oversaturated) background respectively. Similarly, but for blue targets, are represented the cases for both background conditions in **(c)** and **(d)**.

$$\beta = (90-\alpha)(1-\Psi(1...N)) \equiv 90-\theta-\alpha \equiv \arctan(\frac{\Delta S_{B,D}}{\Delta L_{B,D}}), \tag{2.9}$$

$$\Delta S_{D,T} = |S_T - S_D| \equiv |\Delta S_{B,T} - \Delta S_{B,D}| \equiv |\Delta S_{B,T} - (\Delta S_{B,D} \cdot tan(\beta))|, \tag{2.10}$$

$$S_D = \begin{cases} S_T - \Delta S_{D,T} & \text{if } S_T > S_B \\ S_B - \Delta S_{D,T} & \text{otherwise} \end{cases} \tag{2.11}$$

Taking into account these experiments, we wanted to analyze if these search

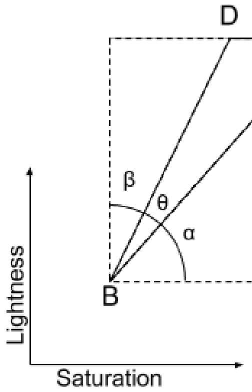
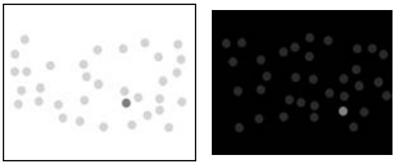


Figure 2.8 – Representation of HSL values for distinct distractors (D), background (B) and search target (T).

asymmetries are present when varying saturation of distractors with respect to a search target. We will see if these differences between distractor and search target are affected by changing background saturation at distinct target and distractor hue (using the HSL color space). A set of stimulus was generated with circular shaped items with a similar displacement to Rosenholtz experiment. Stimulus was sorrounded with a vertical padding equal to the presented background in order to prevent monitor-related luminance gradients. Contrast values can be calculated according to the saturation differences between the search target (T) and distractors (D). Two background (B) conditions were defined, corresponding to Grey (achromatic and unsaturated), and Red (chromatic and oversaturated) colors. At isoluminant ($L_{D,T}$=0.75) and isohue conditions ($H_{D,T}$=0 º for red and $H_{D,T}$=240 º for blue distractors), a representative measure of color contrast between the target and distractors can be computed. This measure was named θ, being the angle between the search target and distractors, with the background as the vertex of the intersection. Same trigonometrical properties can apply using the same diagram plotting B,T and D relationships at distinct quadrants (acknowledging that in our case T is oversaturated for both conditions).

In Equation 2.6, is represented the absolute difference in lightness and saturation between two distinct conditions. In Figure 2.8 there are the angles that comprise the saturation and lightness contrast between our stimulus objects. Each of these angles represent respectively to the triangles formed by B-T (α), B-D (β) and D-T (θ), being α constant for constant background and target (Equations 2.7, 2.8 and 2.9). Most importantly, θ represents the angle comprising the available contrast between the distractor and the target. Given these angle calculations, it is possible to represent the $\Delta S_{D,T}$ as the absolute saturation difference between D and T (Equation 2.10), calculated by the parametrization of β using our psychometric function $\Psi(1...N)$. That absolute saturation difference will define the criterion for our distractor saturation S_D as shown in Equation 2.11. There were generated 4 experimental conditions corresponding to unsaturated and saturated background and red or blue hue. The value of θ is equivalent for saturated and unsaturated background, corresponding to values of $0, 9, 18, 35, 44$ and $53\,^{o}$, producing saturation differences ($\Delta S_{D,T}$) of $0, 0.121, 0.246, 0.528, 0.728$ and 1.

Brightness Contrast (10) According to previous studies, searching a bright target is harder as luminance of distractors increase [220, 238, 300], with a distinct response with respect to chromatic stimuli [210]. Conversely, salience increases for a dark target when luminance of distractors is increased. It was parametrized as the lightness contrast and stimuli was modeled using the HSL color space, considering an achromatic (unsaturated) and isohue relationship between search target, distractors and background, using the same type of stimuli as in Color Contrast (9). Here the target is gray ($L_T = 0.5$) and background is bright ($L_B = 1$) or dark ($L_B = 0$). In order to parametrize the contrast for this stimuli, we used the absolute lightness difference between search target and distractors. In Figure 2.9 theta value is 0 ^{o} for all cases, at 0 saturation, the lighness axis is parametrized.



(a) (b)

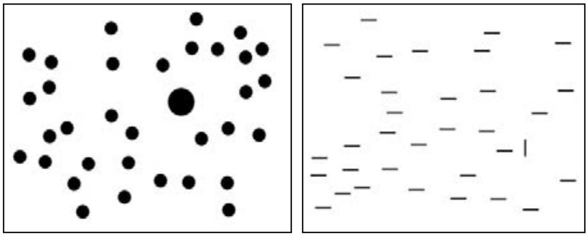
Figure 2.9 – Examples of distinct background conditions, **(a)** lighter background and distractors at L_D=0.66. **(b)** Dark background with distractors at L_D=0.17. For both conditions, the lightness of the search target is grey (L_T=0.5). Absolute contrast for these cases is $\Delta L_{D,T}$=0.33.

$$\Delta L_{D,T} = |L_D - L_T| = \begin{cases} \Delta L_{B,T}(1 - \Psi(1...N)) & \text{if } L_T > L_B \\ \Delta L_{B,T}(\Psi(1...N)) & \text{otherwise} \end{cases} \tag{2.12}$$

$$L_D = \begin{cases} L_T - \Delta L_{D,T} & \text{if } L_T > L_B \\ L_B - \Delta L_{D,T} & \text{otherwise} \end{cases} \tag{2.13}$$

Lightness differences Equation 2.12 are calculated by $\Delta L_{D,T}$, corresponding to the absolute difference between target and background lightness ($|\Delta L_{B,T}|$) and adjusted by our psychometric function $\Psi(1...N)$ with distinct distractor lightness values of L_D, depending on the absolute background lightness with respect the target.

Size contrast (11) Dissimilarities in size of objects tend to drive increase or decrease search speed when detecting and discriminating salient regions [112, 248, 274, 310]. Here is presented size similarity between symmetric objects (circles, without loss of generality). Each stimuli was generated with a set of 34 objects randomly around the scene (Figure 2.10**(a)**). The search target has a distinct size with respect to the distractors, with both cases of smaller and bigger sizes with $\Psi(1...N, 1.25, 5)$=[1.25, 1.67, 2.08, 2.5, 3.34, 4.17, 5] deg, being the size as the parameter that defines the similarity contrast for this case, corresponding to a scaling factor of 0.5 to 2 with respect to the baseline (2.5 deg).

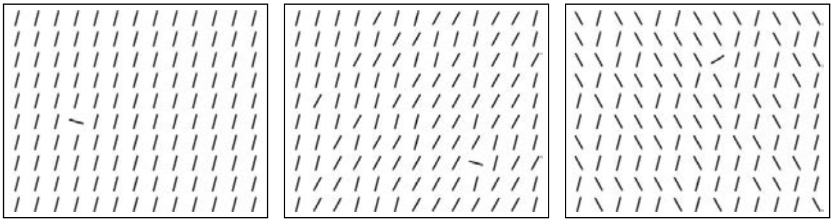


(a) (b)

Figure 2.10 – Examples for salient targets with dissimilar size **(a)** and orientation **(b)**. For **(a)**, the search target has a diameter of 5 deg (a factor of 2 with respect to the rest of distractors). For **(b)**, the orientation of the target is 90 º with distractors at 0 º, forming an orientation contrast of ΔΦ=90 º.

Orientation contrast (12) For this setting, varying angle of objects is found to increase search efficiency when angle contrast is increased [86, 159, 216, 217]. A set of 34 bars were randomly displaced around the scene and oriented at 0 º, in which the search target is an equally-shaped bar oriented at a distinct angle (Figure 2.10(**b**)). Angle contrast between the search target and the set of distractors was $\Delta\Phi(1,0)$=[0, 10, 20, 30, 42, 56, 90]° Equation 2.4.

Distractor Heterogeneity (13) Previous design was to evaluate orientation similarity, given a unique orientation for non-target distractors. Here is presented the phenomenon of distractor heterogeneity. When several sets of distractors are dissimilar with respect to the search target, mutual information between the target and distractors is said to be heterogeneous. In the heterogeneous case, search efficiency is lower, in other terms, target search is harder [17, 86, 101, 216, 267, 319, 359]. Distractor orientation heterogeneity, however, can be represented through distinct configurations, either if the set of distractors are tilted to the same direction or towards distinct directions. In this experiment, there is an array of bars oriented at 75° (with a slope of 15° with respect to the vertical quadrant). From two different sets of distractors, are defined three conditions according to the distinct orientation configurations: homogeneous, tilted-right and flanking (Figure 2.11). For the case of homogeneous distractors, both set of distractors have a unique angle contrast with respect to the target bar. For the case of tilted-right, both set of distractors have an angle tilt of 15 and 30°, $\phi(1,90,15,30)$. For the case of flanking, both sets of distractors have an angle tilt of 15° and −30° respectively, having both positive and negative tilt with respect the search target, $\phi(1,90,15,-30)$.



(a) (b) (c)

Figure 2.11 – Examples of distinct distractor angle configurations, corresponding to **(a)** Homogeneous, **(b)** Tilted-right and **(c)** Flanking.

$$\phi(v,a,c_1,c_2) = \{\Phi(v,a,c_1),\Phi(v,a,c_2)\}, \tag{2.14}$$

$$\Delta\phi(v,a,c_1,c_2) = \{\Delta\Phi(v,a,c_1),\Delta\Phi(v,a,c_2)\}. \tag{2.15}$$

For this type of stimuli, is defined the angle contrast from the search target to two set of distractors, represented as contrasts from the first set "c_1" and the second set "c_2" in Equation 2.15, being the maximum angle between distractors and search target (considering that bars have angle values on two quadrants for each case) as 90 º. Our target angle will have values taken from $\Phi(1,90)$ being parameterized with contrast values ranging from 0 to 90 º, in order to reveal higher angle contrast values, as heterogeneous distractors are harder to be identified.

Distractor Linearity (14) Orientation collinearity facilitates visual guidance when orientation of target differs from its neighbors, making search efficient and in parallel [216, 217, 359]. Visual guidance is induced by orientation linearity given an array of bars as defined from the previous stimulus type. Each bar has been oriented with a specific angle, creating a nonlinear pattern for the whole search array (Figure 2.12). A linear case has also been presented to compare the conspicuity baseline from the other cases.

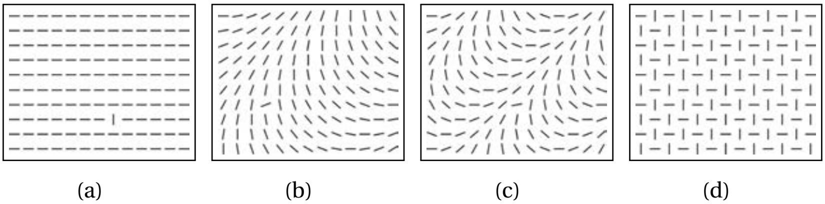


Figure 2.12 – Examples of elicited guidance according to distinct linearity of the distractors. In **(a)**, distractors are set using the same angle contrast of 90 º with respect to the target. Conversely, in **(b)**,**(c)** and **(d)**, nonlinear patterns are set at an accumulative slope of 10, 20 and 90 º respectively.

$$\varphi(u,row,col) = u \cdot row + u \cdot col, \tag{2.16}$$

$$\Delta\varphi(v,u,row,col) = \Phi(v,0) + \varphi(u,row,col). \tag{2.17}$$

Angle contrast is calculated as the orientation difference from the corresponding value of nonlinearity pattern "u" at a certain position on the array "row, col" with a maximum angle contrast with respect to search target (Equations 2.16 and 2.17).

Distractor Categorization (15) Visual search for an oriented bar can be inefficient with distractors at 2 different orientations [101, 359]. However, it was found that not all orientations present on an image are equally coded in pre-attentive vision, different configurations of the orientations of the target and the distractions lead to different discriminability [318]. Some of these orientation configurations were categorized as "steep", in which search target is identified more efficiently. Other categories of heterogeneous distractors presented harder target search and were dependent on set size. The three categories were modeled, corresponding to "steep", "steepest", "steep-right" as defined by Wolfe et. al. [361]. By considering the same orientation contrast between the two sets of angles, target orientation was parametrized in order to reveal at which orientation contrast is the target to both types of distractors that form these categories. We have modeled these three orientation configurations for each distractor pair, corresponding here to −50,50 º for steep, −30,70 º for steepest, 20,80 º for steep-right (Figure 2.13). There was the same amount of distractors for each condition as shown for search on Distractor Heterogeneity (13) and Distractor Linearity (14) in order to uniquely analyze orientation contrast and preserving similar stimulus type conditions. As shown in section Distractor Heterogeneity (13), the orientation values for each set are computed with Equation 2.14 and the contrast with respect to the target as Equation 2.15. The maximum orientation contrast was calculated for all conditions at 40 º ($v = 90/40$) considering the interference of bar orientation contrast in all quadrants (between the target and both distractor orientations). Target angle had psychometric values of $\phi(v, 90, -50, 50)$ for steep, $\phi(v, 90, -30, 70)$ for steepest and $\phi(v, 90, 20, 80)$ for steep-right.

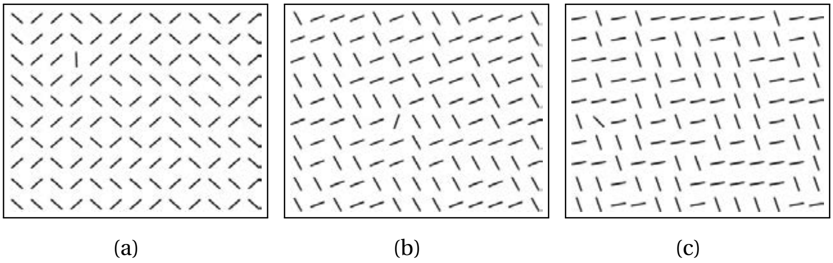


Figure 2.13 – Examples of distinct distractor angle configurations, corresponding to **(a)** Steep, **(b)** Steepest and **(c)** Steep-right.

2.4 Methods and Procedures

Participants

Thirty four subjects (11 female and 23 male) with normal or corrected-to-normal vision took part in this experiment. Most participants were postdoc scholars and PhD students (aged 21–47 years) from non-related fields of study. No economic compensation for the experiments was given. Participants were allowed to wait until they were comfortable with the eye tracking experimental setup in case they had any kind of visual discomfort in between sessions, and they were allowed to adjust the chair while laying on the chin-rest before the experiment. Participants had to sign a consent form allowing the anonymous usage of the data captured during the experiment.

Apparatus

The set of stimulus was presented on a LCD monitor (Samsung SyncMaster HMAQ935729) of screen size 340x270 m m, a resolution of 1280x1080 px and a refresh rate of 60 Hz. A color calibrator was used (Xrite i1 Display Pro) in order to set a specific luminance for the monitor of 160 $\mathrm{cd\,m^{-2}}$, achieving the CIE Illuminant D65 according to the ISO 3664:2000 standard condition (and recommended by Adobe RGB 1998 CIE and ITU-R BT.500-11) with the whitepoint at x=0.313, y=0.329 and a gamma value of 2.2. The light conditions of the room were set using non-direct adjustable light, measured at 30 lx using a luxmeter (TES1332).

We have used a SMI RED binocular eye tracker with a tracking resolution of <0.1 deg, gaze position accuracy of <0.5 deg and a sampling rate of 50 Hz, set at a distance of 600 m m towards the chin-rest (about 40 pixels per degree of visual angle) and vertically equidistant with respect to the monitor, forming a slope of 19 deg from the horizontal axis. The monitor's screen was at a vertical distance of 195 m m from the table and the observer's point of view was adjusted to be centered towards the screen. Fixation and saccade detection was based on SMI iView X Event Detector software, capturing fixations at a minimum duration time of 80 m s and maximum dispersion threshold value of 2 deg and saccades at a peak velocity threshold of 75 deg/s [275][295, p. 243-247].

Procedure

The experiment was divided in one training and two full sessions. During the training session each participant performed a visual search task with 4 types of stimulus with feature and conjunctive search, combined with present and/or absent search targets, hence to ensure their good performance in the next sessions. The first session had a duration of about 20 minutes and was divided in 8 blocks, each one corresponding to a free-viewing or visual search task. The second session had a duration of about 25 minutes and was divided in 10 blocks, similar to the first session. Each task in a block correspond to a distinct stimulus type (shown in Section 2.3) that was presented in a random order. Stimulus order was also randomized across blocks (to avoid any stimulus-related priming [162]), and the location of target distractors was distinct for each case in order to prevent oculomotor biases.

Participants performed two types of tasks: free-viewing and visual search (Figure 2.14). During free-viewing tasks, they were instructed to freely look at the stimuli during 5000 m s. For the visual search task, they were instructed to look for a specific target previously shown in an instruction slide. In case they could find the target, they had to steer their gaze towards it during a dwell time of 1000 m s (the area of interest was based on the target area with an horizontal and vertical spacing of 1 deg). In case they could not identify the target, they were instructed to press a specific key. Considering that context was distinct for each block (replicating stimulus characteristics from previous studies), we decided to do a template target search task instead of an odd-one-out type of task [13, 311]. Participants had unlimited time for the visual search tasks, in this case for reporting target identification or absence. Transitions between stimuli had a duration of 2000 m s (blank transition without the presence of an onset cue) with a luminance equal to the stimuli in order to preserve participant's luminance and chromatic adaptation.

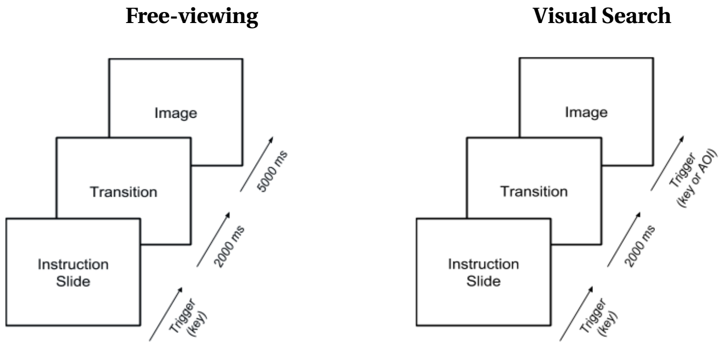


Figure 2.14 – Procedure for the presentation of the stimuli for each task type.

Mean pupil size was recorded to be 2.98 m m diameter for all samples and there was a standard error of 0.18 m m mm between stimulus type, being almost constant throughout the experiment with no significant stimulus-related luminance imbalance.

A 12 point calibration procedure was performed before each session, in which participants were instructed to gaze a red dot moving along different directions. The calibration showed mean deviations for all sessions of σ_x=0.57, σ_y=0.75 deg for the left eye and σ_x=0.56, σ_y=0.82 deg for the right eye. Deviations for each participant's fixation and saccade data were computed using the data of the participant's eye that presented minimum deviation from calibration points in each session. We did a pilot experiment with 4 participants in order to correctly design the visual search procedure, thus, to test the final experimental design for trigger timing and the difficulty of the tasks. That allowed us also to correctly parametrize the variables corresponding to target-distractor contrast where the target was too hard to identify, this parametrization will be shown in the next section.

Data Analysis

In order to get the spatial relevance of participant's eye movements we generated binary maps from fixation coordinates. Fixation density maps are computed with a symmetric Gaussian low-pass filtering (with a window size of [6σ x 6σ]) of the respective binary maps. A value of $\sigma = 1\ deg$ was used, as recommended by LeMeur and Baccino [180], corresponding in our case to 40 pixels. The saliency index (SI) is a measure that relates the energy inside a specific region (that can be manually selected, such as a pop-out region) and the one outside that region.

$$SI(S_t, S_b) = \frac{S_t - S_b}{S_b}. \tag{2.18}$$

We adapted the metric from Spratling's work [300] in order to present positive values as a better representation of the SI (Equation 2.18). The distribution of fixations inside (S_t) and outside (S_b) the area of interest (AOI) will be extracted by cropping the fixation density map using the mask presented on Figure 2.15.

For evaluating the SI for a specific sample, a binary visual mask of the salient region (or AOI) needs to be manually created. Given samples at distinct fixation or saccade number, it is possible to compute gaze-wise SI in order to evaluate the temporal evolution of that measure. Such metric can provide a gold standard of spatial performance in terms of how a region pops out with respect to the rest using fixation density maps from recorded eye movements. In other words, measuring the SI using the fixation density maps is the same as measuring the distribution of fixations that have been recorded inside a particular region of an image. Same parameters of the SI metric are preserved from previous studies [297]. Although other parameters (such as mask area) could be included to better represent the

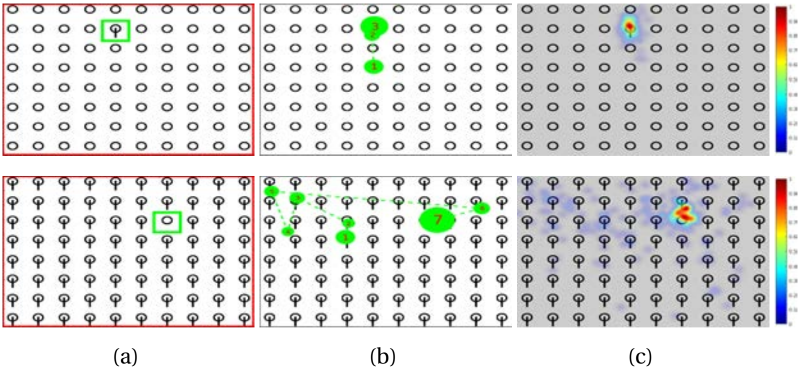


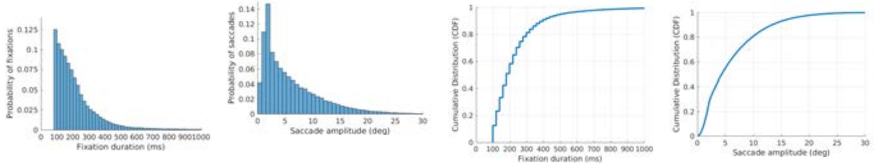
Figure 2.15 – First row shows a grid with 8 × 10 circles, in which one of them becomes salient because of having a superimposed bar. On the second row, the superimposed bar is located instead on the rest of circles (distractors). **(a)** Representation of the mask, corresponding to the AOI of the search target in green (S_t) and the background in red (S_b). **(b)** Example of a scanpath of a single participant, representing each saccade with a green dashed line and each fixation number with a diameter corresponding to its fixation duration. **(c)** Superposed density map from the accumulation of fixations for all participants for such stimuli. The colorbar represents the probability of the density distribution.

data, this could be a metric to be exploited in future studies.

In order to get the performance of participants on salient region localization, we recorded the reaction time (RT) on landing inside the AOI. For the free-viewing tasks, we recorded this RT from the initial fixation until the gaze landed inside the AOI. Once a fixation was outside the AOI, we recorded the time until the gaze returned to the AOI, being in this case produced by inhibition of return (IOR) mechanisms. For the visual search tasks, we recorded in a similar way the first fixation inside the AOI as well as visual discrimination. For this latter case, dwell fixations were pinpointed as being inside the AOI during 1000 ms in order to report identification of search targets. For the cases in which participants could not find the stimulus target, the RT corresponded to key pressing. We used fixation data for reporting target localization on both free-viewing and search tasks and the dwelling method for reporting target identification for visual search tasks. In that way, it is possible to discard non-representative fixations and saccades that could be present by other methods such as key trigger, that could imply spatial and temporal deviations with respect to both visual localization and identification. Given an image where salient regions are known, if the SI and the RT reproduce similar results at distinct tasks, feature contrasts and stimulus types, the SI could provide a way to spatially measure how salient is an object, considering specific regions as pop-out instead of using fixations across the whole scene as ground truth. The usage of eye tracking experiments and regions of interest for calculating localization RT instead of keyboard triggers reveals a more accurate way for evaluating visual attention, as no temporal delays are presented from the time since the participants see the search target to report that they have seen it. That method also allows to prevent them to attend to other regions outside the experimental source over time, such as looking towards the keyboard, which can impair their perceptual adaptability (in terms of light sensitivity and foveation).

2.5 Results

A total of 90,100 fixations were recorded over approximately 30 hours of viewing time. The mean number of fixations per stimulus was $M = 12 \pm 1$, corresponding to $M = 15 \pm 1$ for free-viewing (given from 5000 m s of viewing time) and $M = 11 \pm 1$ for visual search task stimulus (given from the total viewing time until the stimulus trigger, corresponding to target identification). Mean fixation duration was $M = 240 \pm 1 \quad ms$ and it was not presenting significant differences from the two types of tasks. See that both distributions of Fixation Duration (FD) and Saccade Amplitude (SA) (Figure 2.16) were skewed to lower values with their upper and lower quartiles at approximately 100 and 300 m s for FD and 2 and 5 deg of SA. We have also plotted the CDF for both variables and results show that most eye movements (80%) have a FD of less than 300 m s and SA tend to be shorter than 10 deg.



(a) (b) (c) (d)

Figure 2.16 – **(a)** Distribution of Fixation Duration (FD), measured as the absolute fixation time for all samples upon the probability of fixations. **(b)** Distribution of Saccade Amplitude (SA), measured as the absolute euclidean distance between saccade initiation and saccade landing all samples upon the probability of saccades. **(c,d)** Cumulative Distribution Functions for FD and SA.

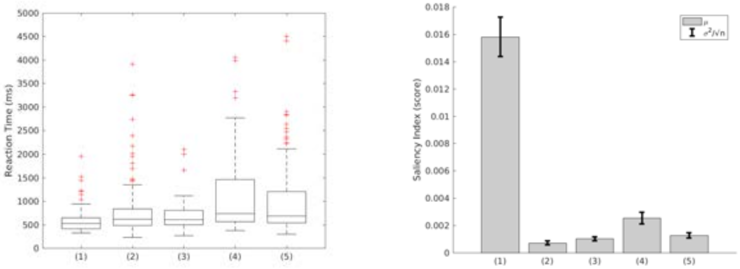
The overall number of fixations was larger for images containing less salient regions, categorized as hard, requiring more fixations for participants in stimulus with less feature contrast. Localization probabilities were calculated, based on the scanpaths in which participants' gaze landed inside the corresponding AOI. Our results report easiest targets more probable to be localized for both free-viewing (p=6.2 × 10^{-4}, Z=3.4, P_{easy}=0.38, P_{hard}=0.30) and visual search tasks (p=6.9 × 10^{-88}, Z=19.9, P_{easy}=0.72, P_{hard}=0.47). After calculating the reaction times for target localization (landing inside the AOI) and identification (reporting presence of target), we discarded samples where $RT > 2\sigma_{RT}$. In that manner we could counteract the impact from oculomotor biases in relation to the localization time with respect objects with approximately that size. As the search targets were smaller for most visual search stimuli, hence less dependent to their respective distance from the stimulus center, we did not discard the respective samples.

2.5.1 Performance upon Feature Type (1st Hypothesis)

RTs for AOI localization are evaluated for each stimulus type and task respectively. Since overall data do not follow a normal distribution (through lilliefors test), Kruskal-Wallis tests were performed in order to evaluate task differences for each contrast difficulty (easy vs hard) as well as differences in RT between distinct type of stimuli. As feature contrasts follow distinct contrast values, we want to test if some features have similarities in RT and their interactions. For each stimulus type, the RT is different given the feature type for both free-viewing (Figure 2.17) and visual search task stimuli (Figure 2.18).

For the former, there were significant differences (p=1.00 × 10^{-10}, $\tilde{\chi}^2$=52.7, $Mdn_{(1)}$=523, $Mdn_{(2)}$=615, $Mdn_{(3)}$=604, $Mdn_{(4)}$=736, $Mdn_{(5)}$=684 m s) between distinct stimulus types RTs, being Corner Angle (1) the fastest stimulus to localize the salient region and Contour Integration (4) the slowest. For the latter, there were significant differences (p=1.11 × 10^{74}, $\tilde{\chi}^2$=372, $Mdn_{(6)}$=782, $Mdn_{(7)}$=742, $Mdn_{(8)}$=942, $Mdn_{(9)}$=892, $Mdn_{(10)}$=593, $Mdn_{(11)}$=787, $Mdn_{(12)}$=622, $Mdn_{(13)}$=606, $Mdn_{(14)}$=676, $Mdn_{(15)}$= 952 m s) on RTs for searching salient regions, showing highest performance for Orientation Contrast (12) and Distractor Categorization (15) the lowest. By computing the saliency index from the density maps across all fixations and the stimulus masks, it is possible to spatially evaluate saliency (in terms of number of fixations inside the window, represented as a heat map isotropically distributed using a Gaussian filter) given from each stimulus types.

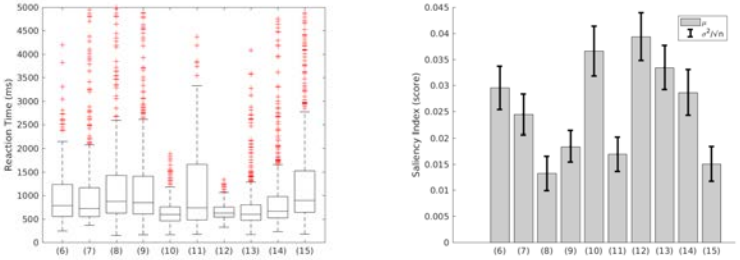
Similarly, there were significant differences on SI depending on stimulus types for free-viewing (p=3.5×10^{-7}, $\tilde{\chi}^2$=36, $Mdn_{(1)}$=1.69×10^{-2}, $Mdn_{(2)}$=6.7×10^{-4}, $Mdn_{(3)}$=1.1× 10^{-3}, $Mdn_{(4)}$=2.2 × 10^{-3}, $Mdn_{(5)}$=1.2 × 10^{-3} and visual search (p=4.9 × 10^{-6}, $\tilde{\chi}^2$=41,



(a)

(b)

Figure 2.17 – Plots for salient region localization time **(a)** and saliency index **(b)** corresponding to stimulus types of Corner Angle (1), Segmentation by Angle (2), Segmentation by Distance (3), Contour Integration (4) and Perceptual Grouping (5)



(a)

(b)

Figure 2.18 – Plots for salient region localization time **(a)** and saliency index **(b)** corresponding to stimulus types of Feature and Conjunctive Search (6), Search Asymmetries (7), Noise/Roughness (8), Color Contrast (9), Brightness Contrast (10), Size Contrast (11), Orientation Contrast (12), Distractor Heterogeneity (13), Distractor Linearity (14) and Distractor Categorization (15).

$Mdn_{(6)}$=32×10^{-3}, $Mdn_{(7)}$=1.4×10^{-2}, $Mdn_{(8)}$=8.0×10^{-3}, $Mdn_{(9)}$=1.8×10^{-2}, $Mdn_{(10)}$=4.0×10^{-2}, $Mdn_{(11)}$=1.6 × 10^{-2}, $Mdn_{(12)}$=4.0 × 10^{-2}, $Mdn_{(13)}$=3.9 × 10^{-2}, $Mdn_{(14)}$=3.1 × 10^{-2}, $Mdn_{(15)}$=13 × 10^{-3}). Stimulus with higher SI for free-viewing task was Corner Angle (1) and the lower was Visual Segmentation (2-3). For the case of visual search task stimuli, most salient targets were on stimulus presented on Size (12) and Orientation (13) contrast and the least ones on Noise/Roughness (8) and Distractor categorization (12) search.

Given the aforementioned results shown for Figures 2.17 and 2.18, RTs were lower (faster) for stimuli with higher SI. The reverse case applies for lower RTs. Target identification (when participats voluntarily report to identify the search target, as explained in Section 2.4) was shown to be slower than target localization (p<1.3 × 10^{-111},Z=−22.4, $Mdn_{localization}$=726, $Mdn_{identification}$=1026 m s), supporting the literature [274] [219] with an absolute mean time difference of M=415 ± 203 ms.

Discussion

We can observe that saliency is induced through varying distinct features of the images. Fixations from participants are shown to localize salient regions significantly with distinct performance depending on feature type and the amount of fixations are distributed or spread distinctively across these regions. These aforementioned observations might be influenced by distinct processing (and correlates) of the visual features in the HVS.

2.5.2 Performance upon Feature Contrast (2nd Hypothesis)

Measures of RTs and SI for salient region localization were computed for each stimulus target-distractor contrast. Overall RT data was not normally distributed, but individual data per stimulus type was normally distributed. Mean RT and error is represented according to the stimulus contrast as well as its mean SI. Spearman's rank correlation tests show that there was a significant negative correlation between RT and SI ($\rho_{RT,SI}$=−.44, $p_{RT,SI}$=2.2 × 10^{-195}), suggesting that SI is a plausible measure for representing saliency on a particular region (higher SI and lower RT implies faster localization speed). In that respect both RT and SI were related to stimulus feature contrast (CT) measurements (shown on Section 2.3 and Figures 2.19,2.20 and 2.21). RT was negatively correlated with respect to CT ($\rho_{CT,RT}$=−.14, $p_{CT,RT}$=7.1 × 10^{-21}). Conversely, SI was correlated with CT $\rho_{CT,SI}$=.05, $p_{CT,SI}$=3.4 × 10^{-3}). These results show that both measurements were satisfying the Weber Law (RT decreasing with higher CT and SI increasing with respect CT). Individual results for correlations between each contrast measurement satisfy for most cases the aforementioned relationships between CT and RT as well as CT and SI, presented in Table 2.3.

We have plotted the relationships between RT and CT as well as for SI and CT in order to see how CT varies localization performance for each stimulus feature type individually (Figures 2.19,2.20 and 2.21). In these figures we can observe (in

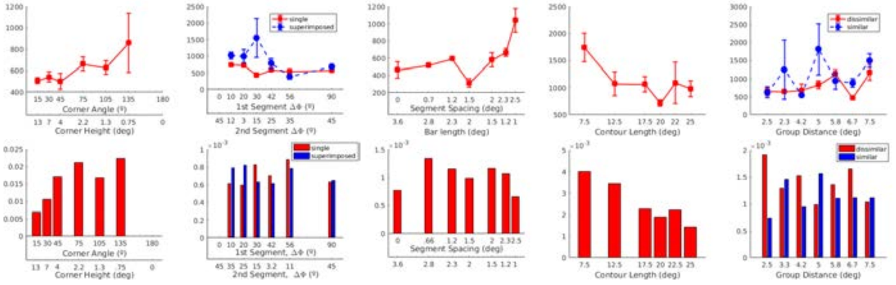
Table 2.3 – Table of correlations between contrast values (3rd column) with Reaction Time (4th column), or with Saliency Index (5th column)

Feature type	Contrast (CT)	ρ_{RT}	ρ_{SI}
(1) Corner Angle	Slope(ž)	.23*	.53*
(2) Segment. Angle	Angle,$\Delta\Phi$(ž)	-.33*	.11
(3) Segment. Spacing	Spacing(deg)	.65*	-.37*
(4) Contour Integration	Length(deg)	-.25*	-.35*
(5) Perc. Grouping	Distance(deg)	.29*	-.06
(6) Feat. & Conj. Search	Set Size(#)	.15*	-.12*
(7) Search Asymmetries	Set Size(#)	-.33*	-.39*
(8) Noise/Roughness	Freq., $1/f^{\beta}$	-.56*	.53*
(9) Color Contrast	Sat.,$\Delta S_{D,T}$	-.57*	.48*
(10) Brightness Contrast	Light.,$\Delta L_{D,T}$	-.41*	.25*
(11) Size Contrast	Size(deg)	-.55*	-.29*
(12) Orientation Contrast	Angle,$\Delta\Phi$(ž)	-.18*	.05
(13) Distr. Heterogeneity	Angle,$\Delta\Phi_{1c}$(ž)	-.04	.17*
(14) Distr. Linearity	Angle,$\Delta\Phi$(ž)	-.07	.01
(15) Distr. Categorization	Angle,$\Delta\Phi_{1c}$(ž)	-.24*	.23*

*: p<.05

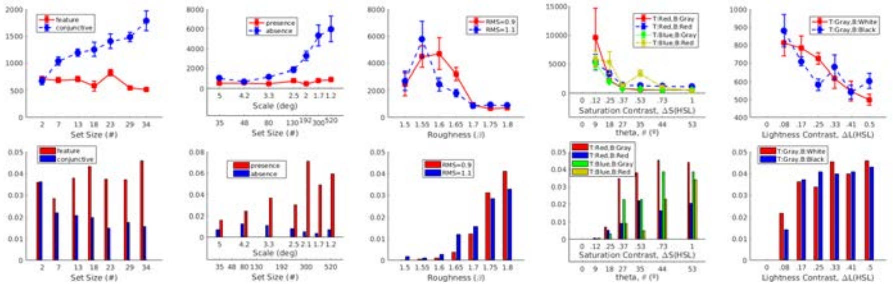
relation to Table 2.3) which feature targets are perceived in parallel or require a serial 'binding' step.

On (1-5) the Weber law applies for stimulus such as Corner Angle, showing slower localization on smoother corners than sharper ones. For Visual Segmentation stimuli, segment localization was faster to be localized when segment angle had a diagonal segment for both single and superimposed segments (due to its own corner angle with respect to other segment bars), being single ones with a trend to be more salient (p=1.2 × 10^{-2}, $\tilde{\chi}^2$=6.3). Segments with 1.5 deg of segment distance and 2.5 deg of bar length showed faster localization rate compared to wider segments. The Weber law applied as well for contour detection, being larger contours faster to be localized. For Perceptual Grouping, similar shape distractors showed slower localization rates as grouping distance is increased (lower proximity), but it was not so evident for dissimilar distractors, being localized faster and with overall higher SI. The Weber law did not apply for this case, suggesting that at a certain proximity distance (about approximately 5.5 deg) participants fixated into several regions, making them similarly salient. SI results on Corner Angle and Contour Integration had positive correlations with respect RT (contradicting the general case). That would be caused by the size of the masks (from stimulus salient objects), which would be higher for higher stimulus contrasts, with decreasing absolute SI (bigger masks would require more fixations when considering the same spatial conditions). In that aspect, SI must be evaluated considering that the size of the



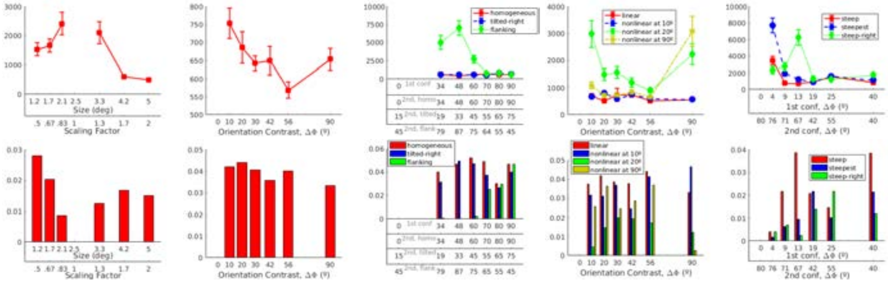
(1) (2) (3) (4) (5)

Figure 2.19 – Plots of Reaction Times (top row) and Saliency Index (bottom row). Spearman's rank correlation tests were performed between RT and SI from each stimulus type and participant individually, corresponding on each case to **Corner Angle (1)**: $\rho_{(1)}$=8.3 × 10^{-2}, $p_{(1)}$=.43, **Visual Segmentation (2,3)**: $\rho_{(2)}$=−.22, $p_{(2)}$=5.6×10^{-3}; $\rho_{(3)}$=−5.7×10^{-4}, $p_{(3)}$=.99, **Contour Integration (4)**: $\rho_{(4)}$=−5.1×10^{-2}, $p_{(4)}$=.61 and **Perceptual Grouping (5)**: $\rho_{(5)}$=−.13, $p_{(5)}$=.12. For this cases, we have discarded samples in which participants had a fixation closer than 5 degrees of eccentricity from the search target, corresponding to the higher visual acuity of the fovea [303][341], as the RT calculation could be impaired by center biases.



(6) (7) (8) (9) (10)

Figure 2.20 – Plots of Reaction Times (top row) and Saliency Index (bottom row). Spearman's rank correlation tests were performed between RT and SI from each stimulus type and participant individually, corresponding on each case to **Feature and Conjunction search (6)**: $\rho_{(6)}$=−.59, $p_{(6)}$=4.6 × 10^{-36}, **Search Asymmetries (7)**: $\rho_{(7)}$=−.45, $p_{(7)}$=3.3 × 10^{-9}, **Noise/Roughness (8)**: $\rho_{(8)}$=−.68, $p_{(8)}$=5.5 × 10^{-33}, **Color Contrast (9)**: $\rho_{(9)}$=−.69, $p_{(9)}$=1.5 × 10^{-72} and **Brightness Contrast (10)**: $\rho_{(10)}$=−.51, $p_{(10)}$=3.4 × 10^{-23}.



(11) (12) (13) (14) (15)

Figure 2.21 – Plots of Reaction Times (top row) and Saliency Index (bottom row). Spearman's rank correlation tests were performed between RT and SI from each stimulus type and participant individually, corresponding on each case to **Size Contrast (11)**: $\rho_{(11)}$=−.14, $p_{(11)}$=9.7 × 10^{-2}, **Orientation Contrast (12)**: $\rho_{(12)}$=−.41, $p_{(12)}$=2.7 × 10^{-8}, **Distractor Heterogeneity (13)**: $\rho_{(13)}$=−.57, $p_{(13)}$=3.5 × 10^{-41} and **Distractor Linearity (14)**: $\rho_{(14)}$=−.57, $p_{(14)}$=2.1 × 10^{-53} and **Distractor Categorization (15)**: $\rho_{(15)}$=−.66, $p_{(15)}$=2.9 × 10^{-59}.

mask is constant, which is not the case for Corner Angle (1) and Contour Integration (4). These center biases might be one of the reasons for the Weber law appliance (presenting less agreement on RT and SI continuity upon feature contrast) as endogenous visual guidance can generate higher inter-participant differences. A more continuous slope for RT and SI observed for stimulus feature contrasts could be acquired by using an onset cue and a constant distance between the initial fixation and the stimulus target, but that method could generate oculomotor biases with respect to the possible positions distinct from the center (that could also vary the temporality of the fixations with respect to the center distance). An alternative solution that would partly solve the problem (as distance from the initial fixation and the stimulus target would still not be totally constant) would be to acquire a larger amount of observations at distinct randomized regions for each stimulus contrast and stimulus type [367][368].

Feature search show a faster localization of the target than conjunctive search (Figure 2.20), with an almost constant RT with respect to set size (features processed in parallel). Conjunction search reveal slower localization of stimulus targets (p=2.5 × 10^{-24}, $\tilde{\chi}^2$=104) as we increase distractor number (consequently, features being shown to be processed in a serial manner), likewise with lower SI. Similarly, reporting stimulus absence presented similar response time distributions, presenting conjunctive distractors to be more uncertain for reporting absence (p=1.9 × 10^{-36}, $\tilde{\chi}^2$=159) than feature search ones. Searching a target circle among circle distractors with a superimposed bar show lower performance at increasing scale and set size (p=6.2 × 10^{-33}, $\tilde{\chi}^2$=143), reversely, searching a target circle with a superimposed bar among circles shows more constant performance, revealing that search asymmetries for this case apply. SI also reveals search asymmetries with respect these two types of stimuli, however, the SI is lower for the former case.

The Weber law is present for the case of background roughness, showing a decrease in search performance and SI at low beta values (rougher surfaces). Both conditions of height deviation (σ_{RMS}=0.9, 1.1) present similar performance with both metrics, with a trend of better search efficiency for higher RMS values. When searching a target with higher saturation contrast with respect distractors, both search performance and SI is higher than with lower saturation contrasts. Background conditions present a trend to drive search asymmetries, showing faster localization RTs and SI for unsaturated backgrounds. In that aspect, achromatic backgrounds presented faster localization for both red (p=3.8 × 10^{-9}) and blue distractors (p=2.1 × 10^{-4}). SI is shown to be higher for red hue in contrast to blue hue for search target and distractors. Lightness contrast also conforms with the Weber law, similarly to saturation contrast but with higher overall performance. Lighter backgrounds with darker search targets present a trend to have higher SI with respect to darker backgrounds with lighter salient objects.

Results on size similarity reveal increased search efficiency with respect to size contrast, with a tendency of perceiving bigger objects as more salient than smaller ones for both localization time and saliency index as in Figure 2.21. Similarity on orientation also shows increased search efficiency with respect angle contrast, with diagonal angles localized faster than vertical or horizontal ones. Orientation contrast has been found to have high search efficiency, specially with diagonal angles and homogeneous angle organization for distractors. In contrast, heterogeneous set of distractor angles present a lower search efficiency with respect to the homogeneous ones. Homogeneous distractors were significantly localized faster for heterogeneity at distinct angle quadrant configurations (flanking) p=1.0 × 10^{-9} but not for heterogeneous distractors with angle configurations at the same quadrant (tilted-right) p=.63. Another distinct type of orientation-related guidance is distractor linearity, presenting differences depending on each slope condition (p=1.3 × 10^{-35}, $\tilde{\chi}^2$=165). Non-linear orientation patterns at a slope increment of 20 º present lower search efficiency than the ones at 10 º and 90 º. The latter case presents a slightly lower search efficiency at vertical or horizontal orientations due to its similar orientation interactions between the target and one of the distractor sets. Results suggest that both the amount of distractor sets and each of their orientation contrasts with respect to search target might be the source of overall distinctiveness for non-linear orientation patterns. Results related with orientation pattern categorization report overall higher SI and a trend for faster localization rate for steep orientation organization than steepest (p=8.4 × 10^{-2}) and significant with respect to steep-right (p=1.2 × 10^{-5}), confirming that search asymmetries apply for this case considering that three conditions possess the same orientation contrast between the two distractor sets.

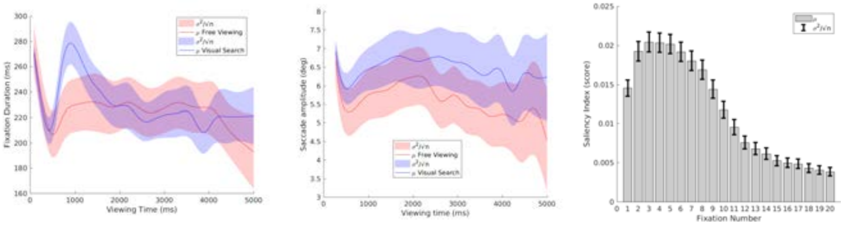
Discussion

In this study we show that feature contrast is correlated to saliency using distinct measures and feature types, being saliency higher at higher feature contrasts. By using visual search tasks and synthetic images, there is a better control of exogenous cues by reducing endogenously-dependent guidance. It is about to consider that the SI is a good measure for evaluating saliency for specific areas of interest.

2.5.3 Attention changes nonlinearly over time (3rd Hypothesis)

Values of FD and SA were grouped for each gaze as functions of viewing time. In Figure 2.22**(a,b)**, during the first 1 to 2 seconds, fixations have a larger duration for visual search tasks. For the visual search task, fixations have a duration with a peak at 274 m s during the beginning of the experiment and progressively drop during the end of the stimulus view to 217 after 5000 m s of viewing time. In free-viewing tasks, FD remains stable after the first and second fixation at approximately 202 m s. For the SA on both tasks there is a peak for the first saccade between 6.5 and 7 deg. During the first and second gaze, SA drops to a value between 5.5 and 6 deg and increase during 1 second to amplitudes between approximately 6 and 6.5 deg. During the last gazes, after 2 seconds of viewing time, SA progressively drops during the rest of viewing time. Such behavior occurs similarly for both visual search and free-viewing cases, these patterns might also be related to endogenous factors commented previously. These distinct eye movement patterns might be related to how participants approach targets depending on task priors and show an overview of how relevant is to account for temporal properties when evaluating eye movements.

SI was computed using the density maps across fixation number Figure 2.22**(c)**, it decreases with respect to fixation number, being the first fixations (from the 1st to the 5th) the ones that have higher SI (accounting for fixations inside the salient region). Inhibition of return (IOR) mechanisms might be responsible for the aforementioned effects. IOR was present and we believe that it may have influenced both types of tasks. To know that, mean return saccade time was computed, corresponding to the time spent from the first fixation inside the AOI to the second fixation that returned inside the AOI, which was M=16.6 ± 0.9 × 10^2 ms, corresponding to M=14.1 ± 0.5 × 10^2 ms for Free-Viewing and M=18.6 ± 1.5 × 10^2ms for Visual Search tasks respectively.



(a)

(b)

(c)

Figure 2.22 – **(a)** Temporal evolution (from 0 to 5000 m s) of fixation duration. **(b)** Temporal evolution (from 0 to 5000 m s) of saccade amplitude. For both plots, samples corresponding to free-viewing task fixations and saccades are represented in red and blue for the case of Visual Search. **(c)** Mean saliency index upon fixation number.

Discussion

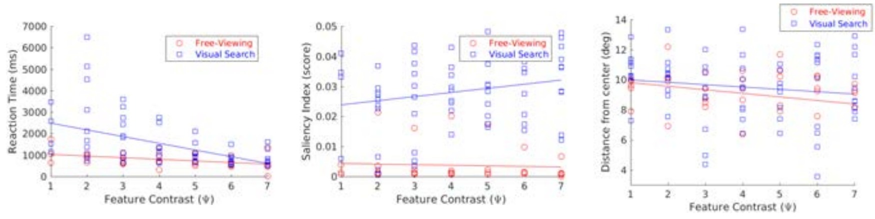
The temporal evolution of fixation and saccade behavior reveal distinct patterns of eye movements upon viewing time , confirming the evidence that visual attention is an active process and its modeling involving temporality requires further investigation. Scanpath prediction could allow the reproduction of the aforementioned effects, regarding in that aspect both bottom-up and top-down processing of visual features that distinctively guide visual attention [39] [165] [66] [202] [1] [3] [344] [356] [351]. In that aspect, as saliency decreases over time, saliency evaluation measures should be done in that line.

2.5.4 Task influences perceived attention (4th Hypothesis)

Distinct eye movement behavior in terms of FD and SA was presented depending on each task type (Section 2.5.3). Task priors also influenced the localization performance in relation to feature contrast.

First, Wilcoxon signed-rank tests were performed to evaluate the amount of fixations between easy and hard targets and was found to be lower for easy than for hard targets in the visual search task (p=2.1 × 10 − 147, Z=−26, Mdn_{easy}=4, Mdn_{hard}=7), but there was no difference for the case of the free-viewing task (p=.069, Z=−0.1, Mdn_{easy}=15, Mdn_{hard}=16). There were differences in FD between the easy and hard targets for visual search (p=3.6 × 10^{-36}, Z=13, Mdn_{easy}=199, Mdn_{hard}=179 ms), but it was not occurring for free-viewing tasks (p=.57, Z=.57, Mdn_{easy}=199, Mdn_{hard}=199 ms). Same phenomena was presented for SA, in which there was a significant difference depending on the stimulus contrast difficulty for visual search (p=1.7 × 10^{-56}, Z=−16, Mdn_{easy}=3.9, Mdn_{hard}=4.7 deg) but not for the case of free-viewing (p=.069, Z=−1.8, Mdn_{easy}=3.9, Mdn_{hard}=4.1 deg). These results evidence less dependence from low-level feature contrasts for free-viewing tasks in contrast to visual search tasks, acknowledging that participants are not always exogenously guided to gaze towards salient regions for free-viewing tasks, namely, that endogenous factors are prevailing more in this kind of task, making saliency less accurate spatially and temporally.

Second, we observed the correlations of RT, SI and feature contrast (FC), described in Section 2.5.2. Here we define FC as ψ values for considering a generalized feature contrast, as CT values vary between blocks, but FC values do not. For visual search stimuli, RT was negatively correlated with SI ($\rho_{RT,SI}$=−.59, $p_{RT,SI}$=.00), FC was negatively correlated with RT ($\rho_{FC,FC}$=−.08, $p_{FC,RT}$=6.4 × 10^{-7}) but positively correlated with SI ($\rho_{FC,SI}$=.05, $p_{FC,SI}$=2.4 × 10^{-3}). For free-viewing stimuli, there was a significant negative correlation between RT and SI ($\rho_{RT,SI}$=−.16, $p_{RT,SI}$=1.2 × 10^{-4}), a negative correlation between FC and RT ($\rho_{FC,RT}$=.26, $p_{FC,RT}$=6.5 × 10^{-10}) but the relationship with respect feature contrast and SI was non-significant ($\rho_{FC,SI}$=−.04, $p_{FC,SI}$=.32). Distinct behavior is presented on the regression lines shown in Figure 2.23 the relationships from RT and SI with respect to CT (here represented as a unique contrast value, although calculated for each contrast measurement separately as in Table 2.3) for both tasks.



(a)

(b)

(c)

Figure 2.23 – Scatter plots of Reaction Time **(a)**, Saliency Index **(b)** and Distance from center **(c)** upon feature contrast (Ψ). We represented the mean of each feature type separately and we have plotted the regression line for both tasks.

Discussion

Salient region localization performance varies with respect to feature contrast depending on the task. Fixation duration and saccade amplitude are affected more by stimulus contrast on Visual Search than Free-Viewing tasks. Moreover, the center bias seems to be more present for Free-Viewing tasks. Further analysis of interest would be the evaluation of absolute task differences in localization performance. In that respect, we could present the same stimuli with several observations for each feature contrast and distinct cueing, so that to see the absolute influences from endogenous guidance for each distinct feature type and contrast.

2.5.5 Center biases are endogenous (5th Hypothesis)

The center bias was represented by grouping fixations for all samples and representing the density map shown in Figure 2.25. From such baseline, it is possible to estimate the mean euclidean distance from every fixation to the baseline center (DC). This baseline shows increasing spreadity and area with respect to fixation number and consequently with respect time. In Figure 2.24 there is the DC as a function of viewing time (centroid was computed as a unique point corresponding to the initial fixation baseline). From this plot, we can observe that participants move their eyes away from the center of the stimulus after the first and second fixation, between 10 and 11 deg. After 2 seconds of viewing time, mean distance from baseline center is nearly constant for the visual search case but not for the free-viewing case. For the latter, fixations get become closer to the baseline center showing increasing patterns of center bias for this task, similarly to SI (Figure 2.22), which increases during the first fixations and drops on late fixations.

DC was negatively correlated with FD (ρ=$-$.08, p=3.9 × 10^{-134}) and positively correlated with SA (ρ=.08, p=1.4 × 10^{-148}). Here, short fixations and large saccades might be eye movement patterns highly related to saliency (as being negatively correlated to the center bias). By computing the mean per stimulus for the case of FC, it is possible to compare how DC was affected by feature contrast Figure 2.23c. For Free-Viewing task, DC was significantly negatively correlated with FC (ρ=$-$.13, p=2.6 × 10^{-3}). For Visual Search task, DC was not significantly correlated with FC (ρ=$-$.03, p=.07). Acknowledging that stimulus targets were randomized, feature contrast was decreasing the center bias (increasing DC) more on Visual Search tasks than for Free-Viewing tasks, supporting the literature stated in Section 2.1.5.

We have added in Table 2.4 the correlations between DC and FC for each feature individually. Most cases of singleton search (i.e. 6-15) show a significant negative correlation between DC and FC, meaning, when the feature contrast is higher, the

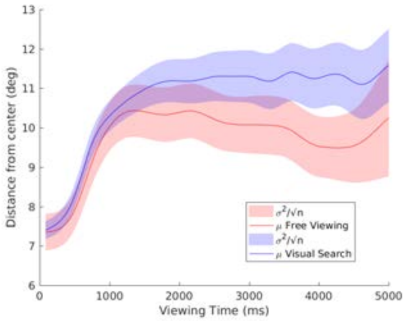
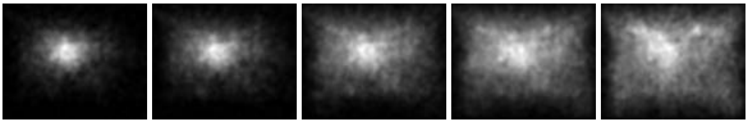


Figure 2.24 – Representation of the center bias as the mean euclidean distance between fixation localization and the baseline center.



Initial **1st** **5th** **10th** **20th**

Figure 2.25 – Representation of the density map for all fixations grouped together across all stimuli.

center bias is lower.

Discussion

Short saccades and large fixation durations are shown to be correlated with eye movement behavior related to the center bias. Temporality of fixations show a non-linear evolution of the center bias, showing more dispersion with respect to viewing time. Moreover, distance from center In that aspect, saliency would not only need to be evaluated by adjusting metric performances using metrics that account for the aforementioned center biases [380][46][33][358][223], but also upon the importance of temporality on fixation and saccade characteristics, by computing each metric upon gaze number on each stimulus fixational data. Thus, saliency metrics should account for feature contrast and minimize the contextual effects in order to accurately reproduce eye movement behavior.

Table 2.4 – Table of correlations between Feature Contrast (FC) with Distance from baseline center (DC)

Feature type	$\rho_{\psi,DC}$
(1) Corner Angle	-.004
(2) Segment. Angle	-.32*
(3) Segment. Spacing	-.16
(4) Contour Integration	.012
(5) Perc. Grouping	-.07
(6) Feat. & Conj. Search	-.34*
(7) Search Asymmetries	.24*
(8) Noise/Roughness	-.29*
(9) Color Contrast	-.12*
(10) Brightness Contrast	.66*
(11) Size Contrast	-.31*
(12) Orientation Contrast	-.18*
(13) Distr. Heterogeneity	.21*
(14) Distr. Linearity	-.28*
(15) Distr. Categorization	-.03

*: p<.05

2.6 General Discussion

Given the presented results, we emphasize that saliency is influenced by a variety of factors when observing eye movement behavior. In this study is presented a dataset considering all the aforementioned factors, by evaluating eye movements for distinct feature types, contrasts, temporality, task and representing the center biases. First, scene context (here defined as different feature types) is known to affect attention with specific performance, significantly determining efficiency of localizing and/or identifying salient regions. Second, saliency measures are shown to be correlated to feature contrast and distinctively depending on feature type. Third, fixation and saccade characteristics are presented to evolve non-linearly over time, making saliency decrease with respect saccade number and/or viewing time. Fourth, visual search tasks show higher performance in comparison to free-viewing on our saliency measurements and they have a higher correlation with respect saliency and feature contrast. Fifth, the central bias is shown to be correlated to short saccades and long fixation durations.

Eye movements are a behavioral output that imply processing of both endogenous and exogenous factors, namely, that have both top-down tuning and bottom-up interactions at different levels of the HVS. Thus, eye movement prediction might require recurrent processing of information from the ventral and dorsal pathways of the HVS, generating a unique representation for eye movement control (visual priority) [172][73][93]. If the unique factor to be evaluated is early saliency, stimulus in which features are processed fast and in parallel would be more relevant when evaluating eye movement prediction (showing less inter-participant differences as a consequence of higher SI), namely, the ones with salient regions that are reflectively selected and separated from the background (with higher contrasts with respect the rest of the scene).

Further considerations

Current literature acknowledges that temporal patterns of saccades have been shown to be fovea-dependent and lately classified as focal and ambient, being ambient fixations responsible for early saccades (sensitive to peripheral signals) and the latter for later saccades (being these ones foveal) [331][232][98][89]. Similarly with saccade latencies, a bimodal latency distribution distinguishes regular from express saccades [277][280][298][332]. We have to acknowledge that the usage of an eye tracker with higher sampling rate (e.g. above 250 Hz) would improve accuracy in this type of experimentation, especially for a possible microsaccadic analysis. Distinct eye movement behavior is presented to be dependent as well for saccade length, pupil dilation and eye vergence [247][205][91][343][249]. All of these factors should be considered in future visual attention modeling considering their relationships with the two-stream hypothesis [333][18][322][292] in order to specify the experimental conditions for a better evaluation of uniquely bottom-up visual attention.

Future work

Future experimentation for low-level feature analysis in eye movements would be to explore covert attention influences varying some of the presented feature contrasts at distinct eccentricities [63][62][60]. Another observation of interest would be the evaluation of task differences in localization performance. In that respect, to present the same stimuli with several observations for each feature contrast and distinct cueing would reveal absolute influences from endogenous guidance. Our study could be extended by analyzing the influence of dynamic scenes on saliency modeling [175][258] using synthetic videos with both static or dynamic camera. In that direction, it would be able to see the interaction between low-level visual features and temporally-variant features. Another remark would be to see the impact of the target template search in comparison to the odd-one out type of tasks, in this case, but for stimuli with similar display conditions but distinct feature type.

Physiological evidence could provide an explanation for the low-level feature processing, including both bottom-up and top-down computations reproducing the presented effects not only spatially but temporally. Computations made by the visual cortex that process these low-level features (in reference to the mechanisms that respond distinctively to color, orientation and spatial sensitivities as well as their interactions) might be responsible for most if not all of the effects presented in this study. Further analysis on mid and high-level features would require further study in terms of their relation to psychophysical effects on eye movements as well as their biological foundations [164].

3 SID4VAM: Synthetic Image Dataset for Visual Attention Modeling

3.1 Objectives

Visual saliency is a term coined on a perceptual basis. According to this principle, a correct modelization of saliency should consider specific experimental conditions upon a visual attention task. The output of such a model can vary for stimulus or task, but must arise as a common behavioral phenomena in order to validate the general hypothesis definition from Treisman, Wolfe, Itti and colleagues [145, 321, 365]. Eye movements have been considered the main behavioral markers of visual attention. But understanding saliency means not only to prove how visual fixations can be predicted, but to simulate which patterns of eye movements are gathered from vision and its sensory signals (here avoiding any top-down influences). This challenge offers eye tracking researchers to consider several experimental issues (with respect contextual, contrast, temporal, oculomotor and task-related biases) when capturing bottom-up attention, largely explained by Borji et al. [41], Bruce et al. [51] and lately by Berga et al. [23]. Computational models advance several ways to predict, to some extent, human visual fixations. However, the limits of the prediction capability of these saliency models arise as a consequence of the validity of the evaluation from eye tracking experimentation. We aim to to provide a new dataset with uniquely synthetic images and a benchmark, studying for each saliency model:

1. How model inspiration and feature processing influences model predictions?

2. How does temporality of fixations affect model predictions?

3. How low-level feature type and contrast influences model's psychophysical measurements?

3.2 Previous and current literature

In order to determine whether an object or a feature attracts attention, initial experimentation was assessing feature discriminability upon display characteristics (e.g. display size, feature contrast...) during visual search tasks [321, 359, 365]. Parallel search occurs when features are processed preattentively, therefore search targets are found efficiently regardless of stimulus properties. Instead, serial search happens when attention is directed to one item at a time, requiring a "binding" process to allow each object to be discriminated. For this case, search time decrease with feature contrast or set size (following the Weber Law [92]).

Table 3.1 – Characteristics of eye tracking datasets

A: Real Images

Dataset	Task	# TS	# PP	PM	DO
Toronto [50]	FV	120	20		✓
MIT1003 [151]	FV	1003	15		✓
NUSEF [252]	FV	758	25		✓
KTH [160]	FV	99	31		✓
MIT300 [150]	FV	300	39		✓
CAT2000 [42]	FV	4000	24		✓

B: Psychophysical Pattern / Synthetic Images

Dataset	Task	# TS	# PP	PM	DO
iLab USC [145]	-	~540	-	✓	
UCL [385]	VS & SG	2784	5	✓	
VAL Harvard [367]	VS	4000	30	✓	
ADA KCL [300]	-	~430	-	✓	
CAT2000$_p$ [42]	FV	100	18		✓
SID4VAM (Ours)	FV & VS	230	34	✓	✓

TS: total number of stimuli, PP: participants, PM: Parametrization, DO: Fixation data is available online, FV: Free-Viewing, VS: Visual Search, SG: visual segmentation

More current studies replicated these experiments by providing real images with parametrization of feature contrast and/or set size (iLab USC, UCL, VAL Hardvard, ADA KCL), combining visual search or visual segmentation tasks, however not providing eye tracking data (Table 3.1B). Rather, current eye movement datasets provide fixations and scanpaths from real scenes during free-viewing tasks. These image datasets are usually composed of real image scenes (Table 3.1A), either from indoor / outdoor scenes (Toronto, MIT1003, MIT300), nature scenes (KTH) or semantically-specific categories such as faces (NUSEF) and several others (CAT2000). A complete list of eye tracking datasets is in Winkler & Subramanian's overview [355]. CAT2000 training subset of "Pattern" images (CAT2000$_p$) provides eye movement data with psychophysical / synthetic image patterns during 5 sec of free-viewing. However, no parametrization of feature contrast nor stimulus properties is given. A synthetic image dataset could provide information of how attention is dependent on feature contrast and other stimulus properties with distinct tasks.

3.3 Dataset

As explained in [Chapter 2], fixations were collected from 34 participants in a dataset of 230 images [1]. Images were displayed in a resolution of 1280 × 1024 px and fixations were captured at about 40 pixels per degree of visual angle using SMI RED binocular eye tracker. The dataset had been splitted in two tasks: Free-Viewing (FV) and Visual Search (VS). For the FV task, participants had to freely look at the image during 5 seconds. On each stimuli there was a salient area of interest (AOI). For the VS task, participants had the instruction to visually locate the AOI, setting the salient region as the different object. For this task, the trigger for prompting the transition to next image was by gazing inside the AOI or pressing a key (for reporting absence of target). We can observe the stimuli generated for both tasks on Figs. 3.1-3.2.

The dataset was divided in 15 blocks, 5 corresponding to FV and 10 to VS. Some of these blocks had distinct subsets of images (due to the alteration of either target or distractor shape, color, configuration and background properties), abling a total of 33 types of stimuli. Each of these blocks was individually generated as a low-level feature category, which had its own type of feature contrast between the salient region and the rest of distractors / background. FV categories were mainly based for analyzing preattentive effects (Fig. 3.1): 1) Corner Salience, 2) Visual Segmentation by Bar Angle, 3) Visual Segmentation by Bar Length, 4) Contour Integration by Bar Continuity and 5) Perceptual Grouping by Distance. VS categories were based on a feature-singleton search stimuli, where there was a unique salient target and a set of distractors and/or altered background (Fig. 3.2). These categories were: 6) Feature and Conjunctive Search, 7) Search Asymmetries, 8) Search in a Rough Surface, 9) Color Search, 10) Brightness Search, 11) Orientation Search, 12) Dissimilar Size Search, 13) Orientation Search with Heterogeneous distractors, 14) Orientation Search with Non-linear patterns, 15) Orientation search with distinct Categorization. Stimuli for SID4VAM's dataset was inspired by previous psychophysical experimentation [300, 359, 385].

Dataset stimuli were manually generated with 7 specific instances of feature contrast (Ψ), corresponding to hard ($\Psi_h = \{1..4\}$) and easy ($\Psi_e = \{5..7\}$) difficulies of finding the salient regions. These feature contrasts had their own parametrization corresponding to the feature differences between the salient target and the rest of

[1] Download dataset: http://www.cvc.uab.es/neurobit/?page_id=53

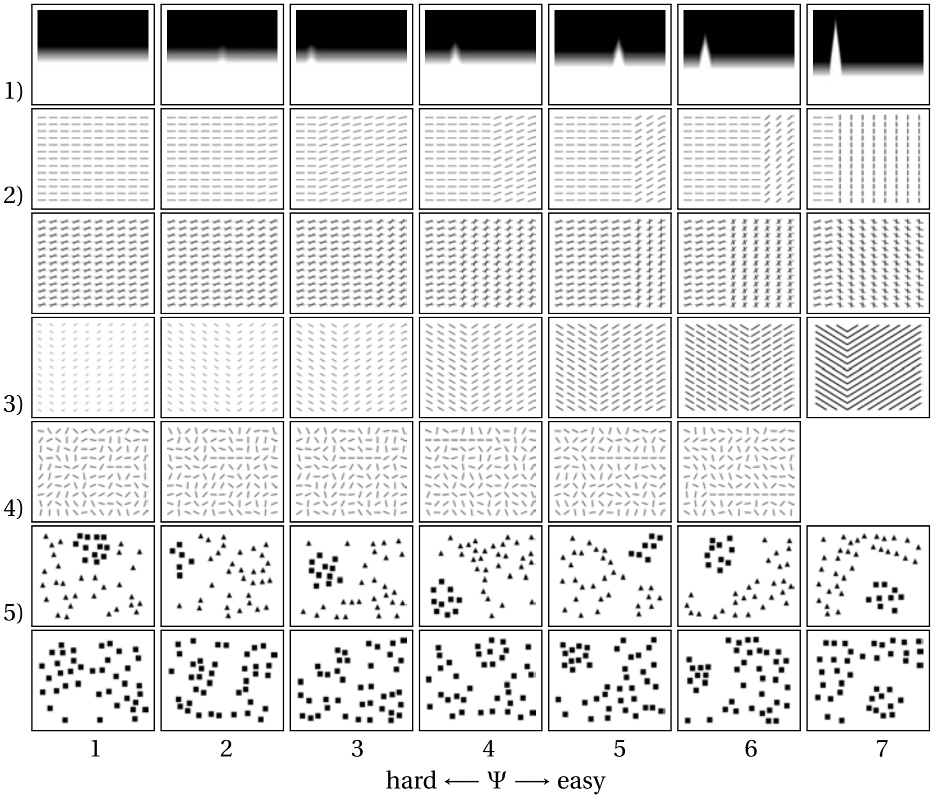


Figure 3.1 – Free-Viewing stimuli: **1)** Corner Angle, **2-3)** Visual Segmentation, **4)** Contour Integration and **5)** Perceptual Grouping

distractors (e.g. differences of target orientation, size, saturation, brightness...) or global effects (e.g. overall distractor scale, shape, background color, background brightness).[2]

[2] Code for generating synthetic stimuli: https://github.com/dberga/sig4vam

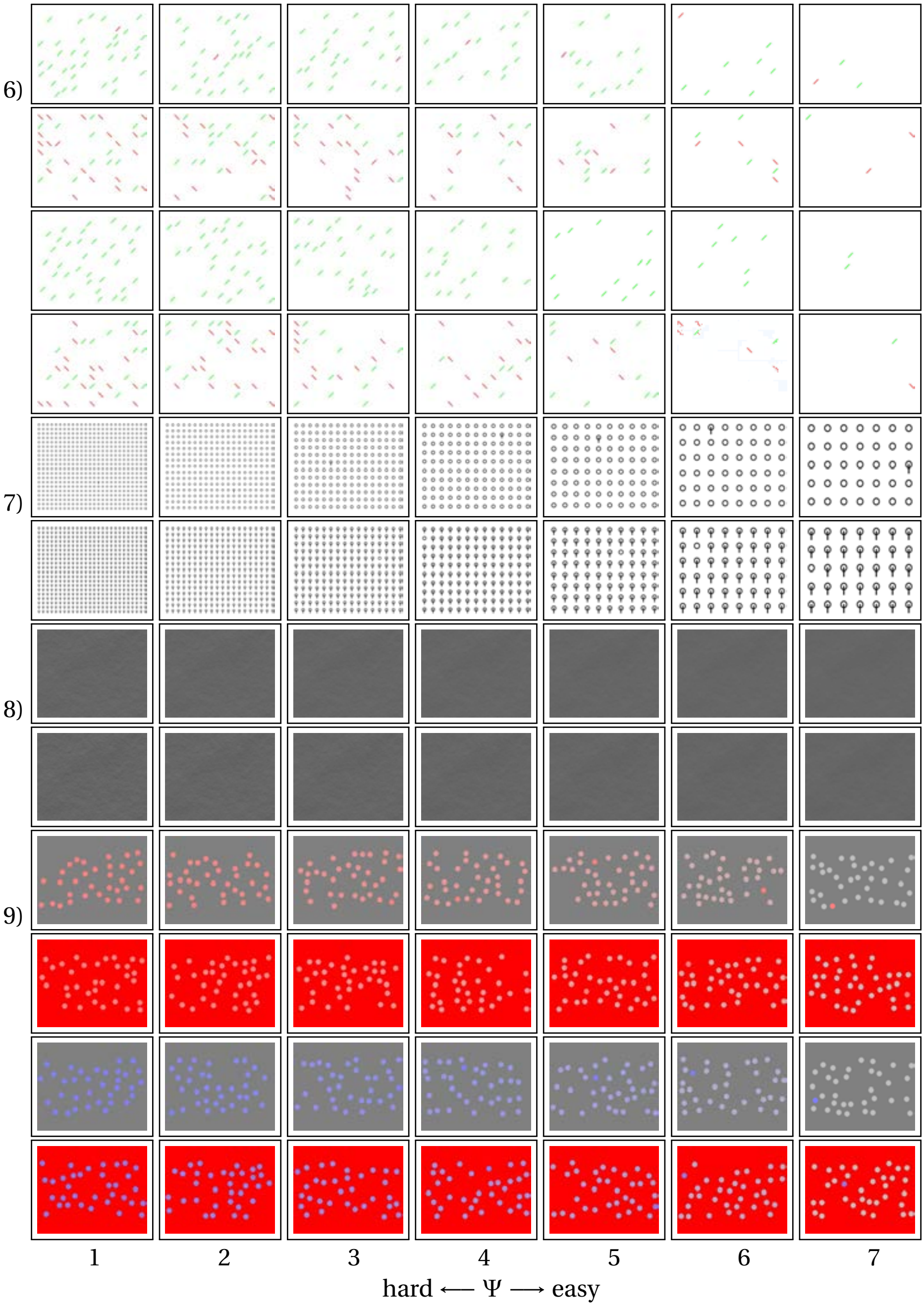
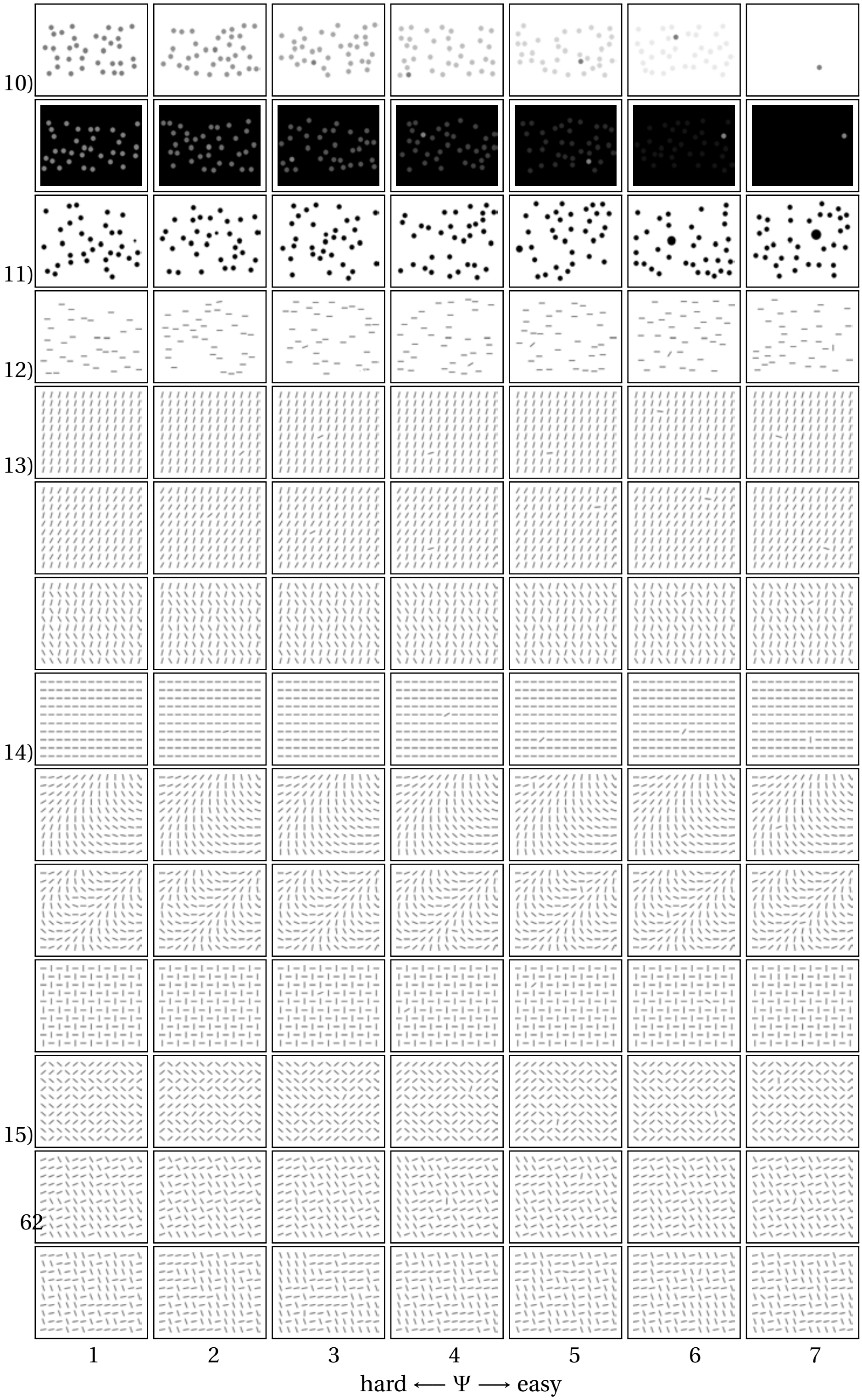


Figure 3.2 – Visual Search stimuli: **6)** Feature and Conjunctive Search, **7)** Search Asymmetries, **8)** Roughness, **9-10)** Color and Brightness contrast, **11)** Size contrast, **12)** Orientation contrast in **13)** Heterogeneous, **14)** Nonlinear and **15)** Categorical search.


10)
11)
12)
13)
14)
15)
1
2
3
4
5
6
7
hard ⟵ Ψ ⟶ easy

3.4 Methods and Procedure

Fixation maps from eye tracking data are generated by distributing each fixation location to a binary map. Fixation density maps are created by convolving a gaussian filter to the fixation maps, this simulates a smoothing caused by the deviations of σ=1 deg given from eye tracking experimentation, recommended by LeMeur & Baccino [180].

Typically, location-based saliency metrics (AUC_{Judd}, AUC_{Borji}, NSS) increase their score fixation locations fall inside (TP) the predicted saliency maps. Conversely, scores decrease fixation locations are not captured by saliency maps (FN) or when saliency maps exist in locations with no present fixations (FP). In distribution-based metrics (CC, SIM, KL), saliency maps score higher when they have higher correlations with respect to fixation density map distributions. We have to point out that shuffled metrics (sAUC, InfoGain) consider FP values when saliency maps coincide with other fixation map locations or a baseline (here, corresponding to the center bias), which are not representative data for saliency prediction. Prediction metrics and its calculations are largely explained by Bylinskii et al. [55]. Our saliency metric scores and pre-processing used for this experimentation have been replicated from the official saliency benchmarking procedure [54]. Psychometric evaluation of saliency predictions has been done with the Saliency Index (SI) [297, 300]([Equation 2.18]. Model evaluations have been divided according to its inspiration and prediction scores have been evaluated with saliency metrics and in psychophysical terms.

3.5 Results on predicting fixations

Previous saliency benchmarks [43, 51, 53, 54, 256] reveal that Deep Learning models such as SALICON, ML-Net SAM-ResNet, SAM-VGG, DeepGazeII or SalGan score highest on both shuffled and unshuffled metrics. In this section we aim to evaluate whether saliency maps that scored highly on fixation prediction do so with a synthetic image dataset and if their inspiration influences on their performance. We present metric scores of saliency map predictions of the whole dataset in Table 3.2 and plots in Fig. 3.3. Saliency metric scores reveal that overall Spectral/Fourier-based saliency models predict better fixations on a synthetic image dataset.

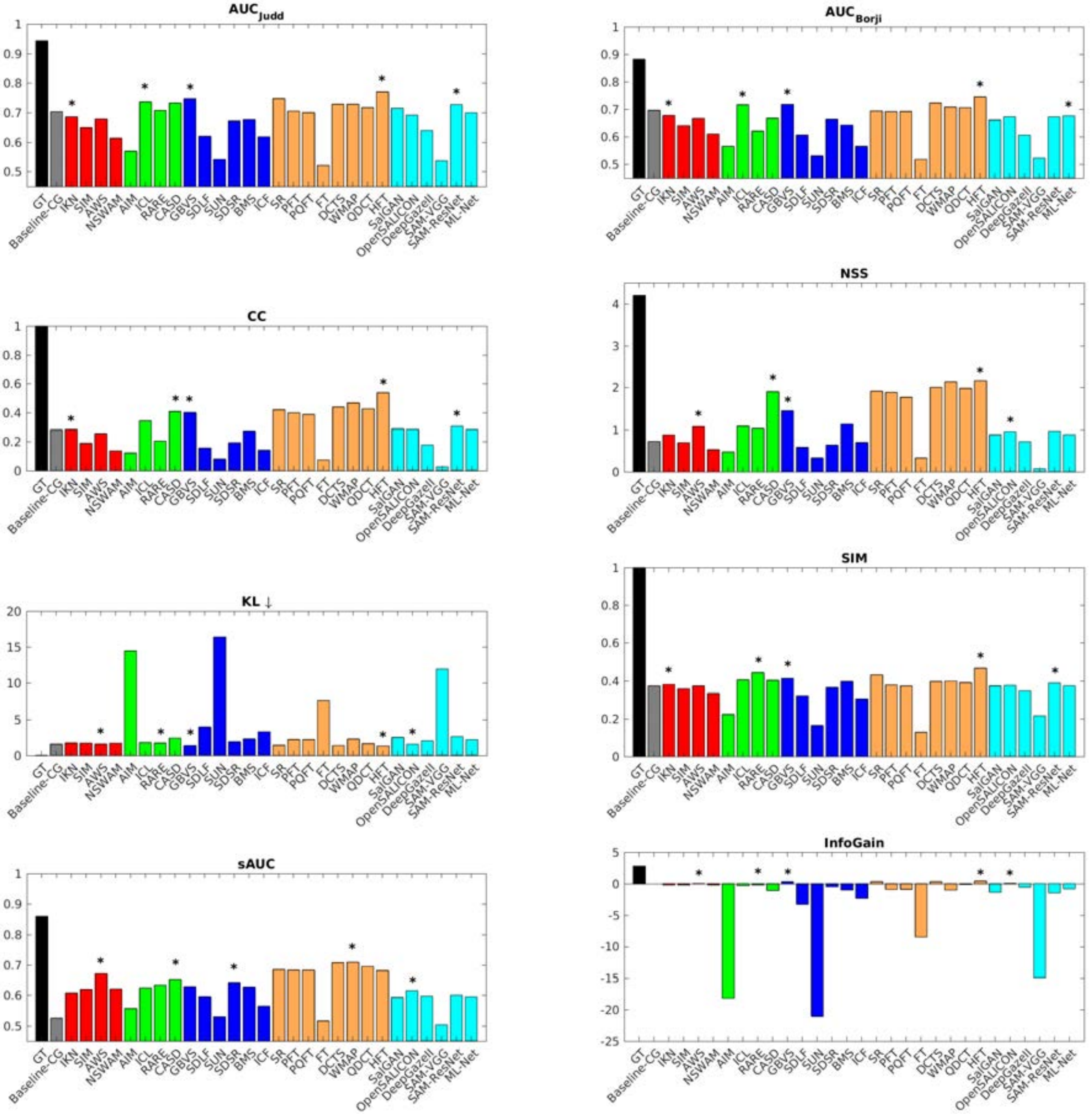


Figure 3.3 – Plots for saliency metric scores for SID4VAM dataset

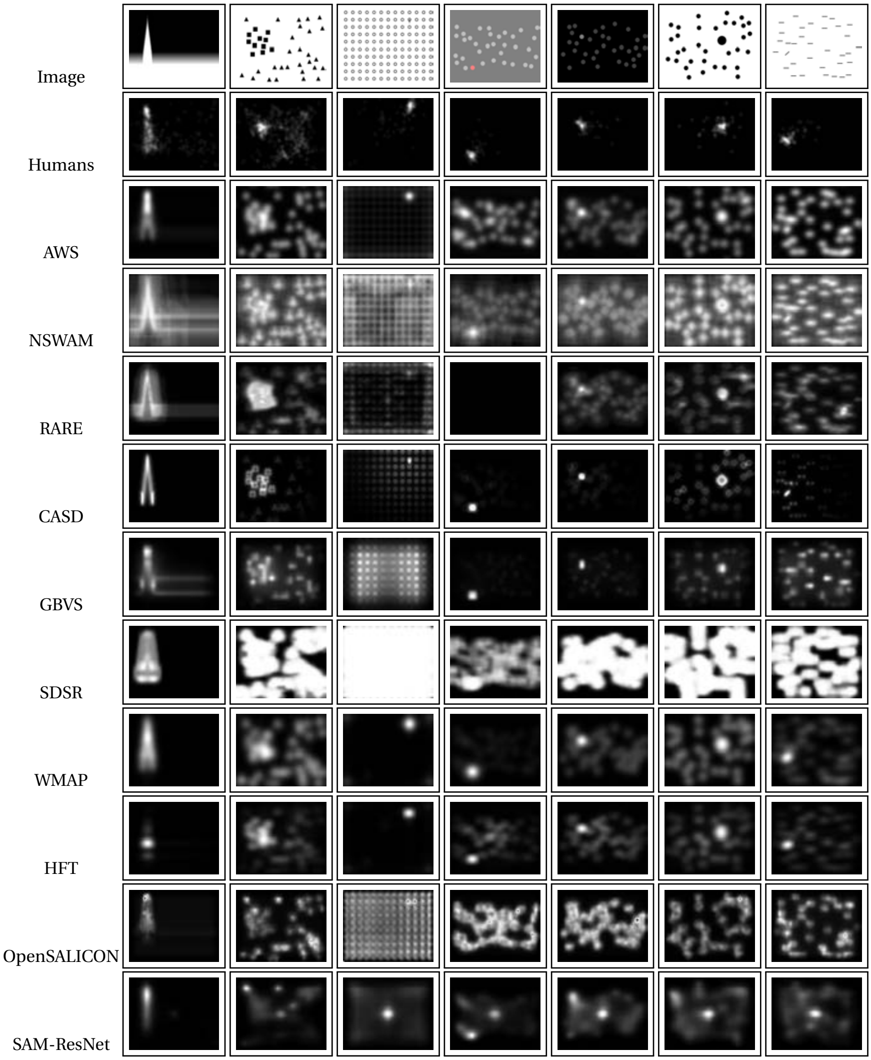


Figure 3.4 – Examples of dataset stimuli and saliency map predictions. Only two models for each inspiration category that presented highest performance with shuffled saliency metric scores (sAUC and InfoGain) are shown.

Models such as HFT and WMAP remarkably outpeform other saliency models. From other model inspirations, AWS score higher than other Cognitive/Biologically-inspired models, GBVS and CASD outperform other Probabilistic/Bayesian and Information-theoretic saliency models respectively. For Deep Learning models, SAM$_{ResNet}$ and OpenSALICON are the ones with highest scores. Although there are present differences in terms of model performances and model inspiration, similarities in model mechanisms can reveal phenomena of increasing and decreasing prediction statistics. This phenomena is present for Spectral/Fourier-based and Cognitive/Biologically-inspired models, withwhom all present similar performance and balanced scores throughout the distinct metric scores. It is to consider that sAUC and InfoGain metrics are more reliable compared to other metrics (which the baseline center gaussian sometimes acquires higher performance than most saliency models). In these terms, models shown on Fig. 4.6 are efficient saliency predictors for this dataset. We can also point out that models which process uniquely local feature conspicuity scored lower on SID4VAM fixation predictions, whereas the ones that processed global conspicuity scored higher. This phenomena might be related with the distinction of foveal (near the fovea) and ambient (away from the fovea) fixations, relative to the fixation order and the spatial locations of fixations [89, 98]. The evaluation of gaze-wise model predictions has been done by grouping fixations of every instance separately. We have plotted results of the $sAUC$ saliency metric for each model (Fig. 3.5) and it is observable that model performance decrease upon fixation number, meaning that saliency is more likely to be predicted during first fixations. For evaluating the temporal relationship between human and model performance ($sAUC$), we have performed Spearman's (ρ) correlation tests for each fixation and it can be observed that IKN, ICL, GBVS, QDCT and ML-Net follow a similar slope as the GT, contrary to the case of the baseline center gaussian.

Table 3.2 – Saliency metric scores for SID4VAM

Model	AUCj	AUCb	CC	NSS	KL	SIM	sAUC	InfoGain
GT	0.943	0.882	1.000	4.204	0.000	1.000	0.860	2.802
Baseline-CG	0.703	0.697	0.281	0.722	1.577	0.372	0.525	-0.189
IKN	0.686	0.678	0.283	0.878	1.748	0.380	0.608	-0.233
SIM	0.650	0.641	0.189	0.694	1.702	0.357	0.619	-0.148
AWS	0.679	0.667	0.255	1.088	1.592	0.373	0.672	0.013
NSWAM	0.614	0.610	0.136	0.529	1.686	0.335	0.622	-0.150
AIM	0.570	0.566	0.122	0.473	14.472	0.224	0.557	-18.182
ICL	0.737	0.717	0.343	1.100	1.788	0.405	0.624	-0.313
RARE	0.707	0.622	0.204	1.046	1.736	0.444	0.633	-0.158
CASD	0.733	0.669	0.408	1.904	2.395	0.403	0.652	-1.046
GBVS	0.747	0.718	0.400	1.464	1.363	0.413	0.628	0.331
SDLF	0.620	0.607	0.156	0.585	3.954	0.322	0.596	-3.244
SUN	0.542	0.532	0.080	0.333	16.408	0.165	0.530	-21.024
SDSR	0.672	0.665	0.192	0.639	1.904	0.365	0.642	-0.467
BMS	0.677	0.643	0.274	1.143	2.306	0.397	0.627	-0.958
ICF	0.618	0.566	0.141	0.700	3.274	0.306	0.564	-2.300
SR	0.748	0.694	0.420	1.916	1.432	0.431	0.685	0.348
PFT	0.705	0.692	0.398	1.885	2.227	0.377	0.684	-0.893
PQFT	0.701	0.693	0.387	1.774	2.197	0.373	0.684	-0.856
FT	0.521	0.518	0.072	0.331	7.552	0.129	0.517	-8.498
DCTS	0.729	0.724	0.439	2.004	1.363	0.396	0.708	0.337
WMAP	0.729	0.709	0.468	2.136	2.283	0.397	**0.709**	-0.981
QDCT	0.717	0.706	0.425	1.986	1.677	0.391	0.695	-0.105
HFT	**0.771**	**0.746**	**0.538**	**2.161**	**1.295**	**0.467**	0.682	**0.448**
SalGAN	0.715	0.662	0.287	0.883	2.506	0.373	0.593	-1.350
OpenSALICON	0.692	0.673	0.284	0.956	1.549	0.375	0.615	0.052
DeepGazeII	0.639	0.606	0.176	0.714	2.023	0.346	0.597	-0.587
SAM-VGG	0.537	0.523	0.026	0.070	11.947	0.216	0.503	-14.954
SAM-ResNet	0.727	0.673	0.305	0.967	2.610	0.388	0.600	-1.475
ML-Net	0.700	0.676	0.283	0.883	2.169	0.373	0.595	-0.837

Cognitive/Biological , Information-Theoretic , Probabilistic , Fourier/Spectral ,
Machine/Deep Learning

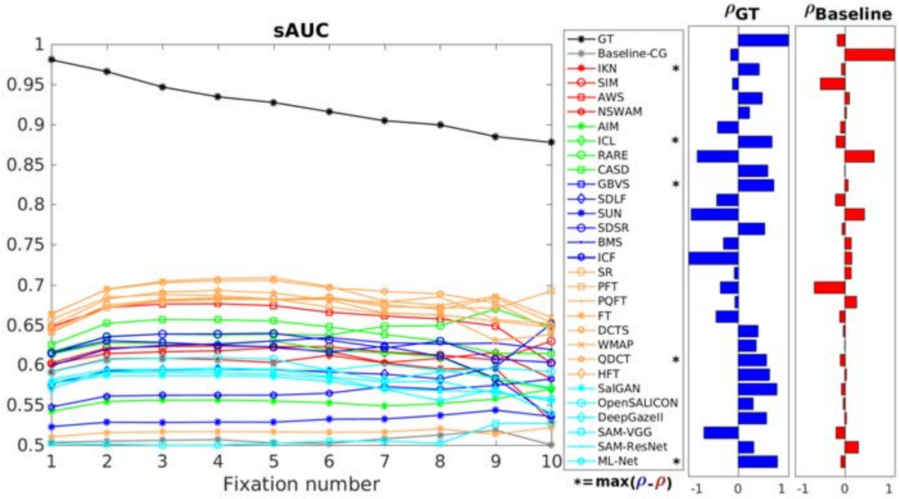


Figure 3.5 – sAUC gaze-wise prediction scores.

3.6 Results on psychophysical consistency

Previous studies [23, 41, 51] found that several factors such as feature type, feature contrast, task, temporality of fixations and the center bias alternatively contribute to eye movement guidance. The HVS has specific contrast sensitivity to each stimulus feature, so that saliency models should adapt in the same way in order to be plausible in psychometric parameters. Here we will show how saliency prediction varies significantly upon feature contrast and the type of low-level features found in images. In Fig. 3.6a is found that saliency models increase SI with feature contrast "Ψ" following the distribution of human fixations. Most prediction SI scores show a higher slope with easy targets (salient objects with higher contrast with respect the rest, when Ψ > 4), being CASD and HFT the models that have higher SI at higher contrasts.

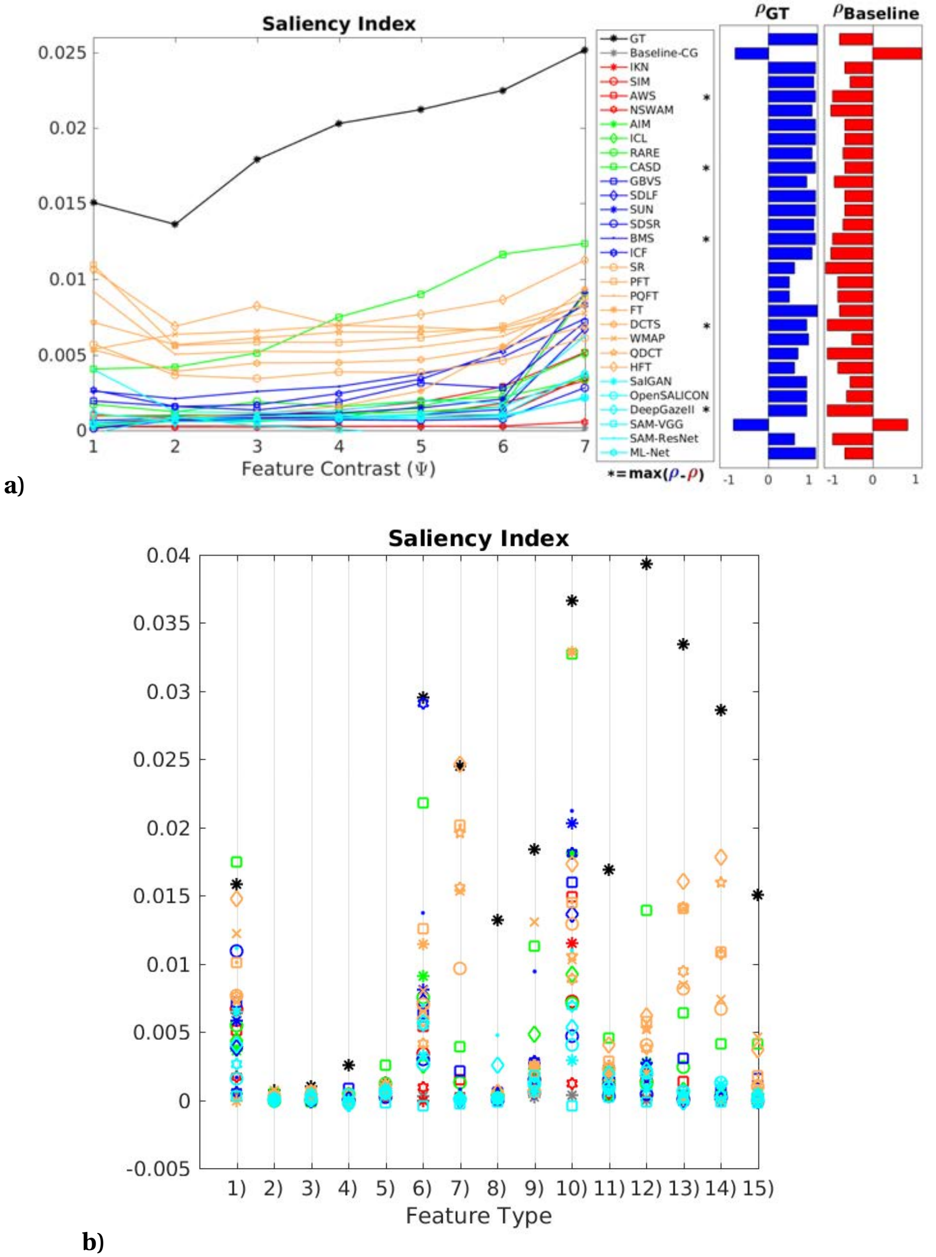
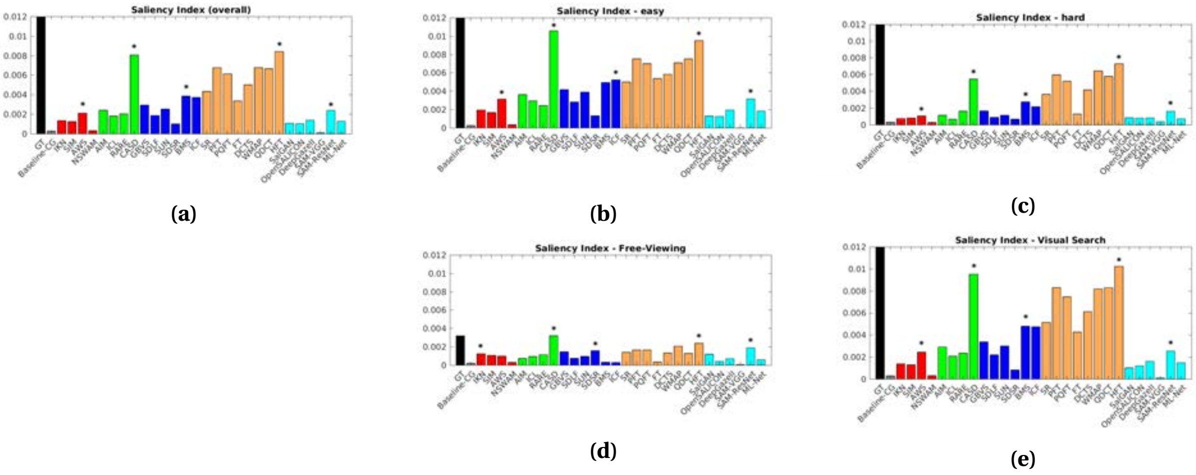


Figure 3.6 – Results of Saliency Index of model predictions upon Feature Contrast **(a)** and Feature Type **(b)**.

Contextual influences (here represented as distinct low-level features that appear in the image) contribute distinctively on saliency induced from objects that appear on the scene [140]. We suggest that not only the semantic content that appears on the scene affects saliency but the feature characteristics do significantly impact how salient objects are. This phenomena is observable in Fig. 3.6b and occurs for both human fixations and model predictions, specifically with highest SI for human fixations in 1) Corner Salience, 6) Feature and Conjunctive Search, 7) Search Asymmetries, 10) Brightness Search, 12) Dissimilar Size Search and 13) Orientation Search with Heterogeneous distractors. HFT and CASD have highest SI when GT is higher (when human fixations are more probable to fall inside the AOI), even outperforming GT probabilities for the cases of 1) and 7). We show in Fig. 3.7a that overall Saliency Index of most saliency models is distinct when we vary the type of feature contrast (easy vs hard) and the performed stimulus task (free-viewing vs visual search). Spectral/Fourier based models outperform other saliency models also in SI metric. Similarly with saliency metrics shown on previous subsection, AWS, CASD, BMS, HFT and SAM-ResNet are the most efficient models for each model inspiration category respectively. It is observable in Fig. 3.7b that saliency models have higher performance for easy targets, with increased overall model performance differences with respect hard targets (Fig. 3.7c). Similarly, visual search targets show lower difficulty (higher SI) to find predicted fixations inside the AOI than the free-viewing cases (Fig. 3.7d-e).


Saliency Index (overall)

(a)

Saliency Index - easy

(b)

Saliency Index - hard

(c)

Saliency Index - Free-Viewing

(d)

Saliency Index - Visual Search

(e)

Figure 3.7 – Results of Saliency Index metric scores from dataset model predictions **(a)**, for easy/hard difficulties **(b-c)** and Free-Viewing/Visual Search tasks **(d-e)**.

Also distinct SI curves upon feature contrast are reported, revealing that contrast sensitivies are distinct for each low-level feature. Spearman's correlation tests on Fig. 3.6b show which models correlate with human performance over feature contrast and which one do so with the baseline (designating higher center biases). These results show that models such as AWS, CASD, BMS, DCTS or DeepGazeII highly correlate with human contrast sensitivities and do not correlate with the baseline center gaussian. Separate results are shown for each feature type in Fig. 3.8, showing distinct performances of predicted saliency maps. Matching human contrast sensitivities on low-level visual features would be an interesting point of view to make future saliency models accurately predict saliency as well as to better understand how the HVS processes visual scenes.

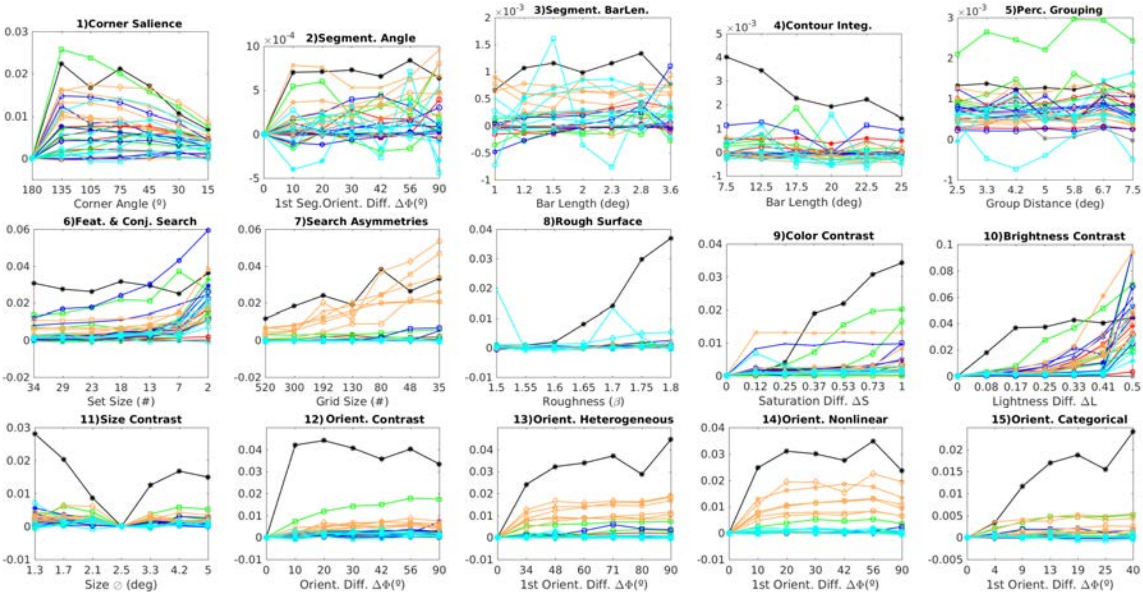


Figure 3.8 – Plots of Saliency Index from saliency models upon feature contrast for each feature type (mean of all block subcategories for each contrast). **1st row:** Free-Viewing stimuli. **2nd-3rd row:** Visual Search stimuli.

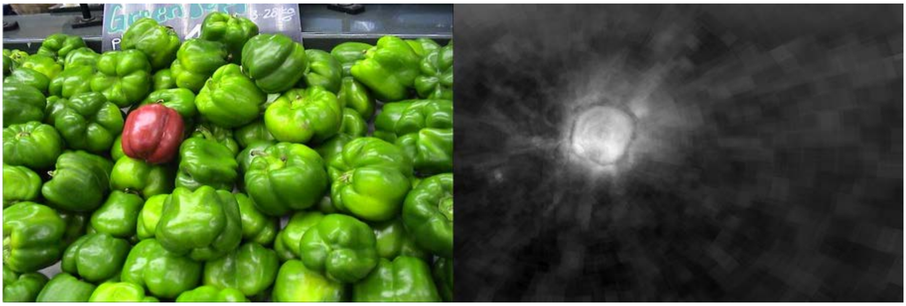
3.7 Discussion

Previous saliency benchmarks show that saliency is efficiently predicted with latest Deep Learning saliency models. This is not the case with synthetic images. A possible reason for this is that Machine/Deep Learning models are trained uniquely with datasets that contain high-level features (i.e. indoor and outdoor real images with animated and unanimated objects), thus, overfitting this type of contextual information. Another possibility is that we randomly determined where salient objects are, making the center bias affect less to our experimentation. With this benchmark we can evaluate how salient is a particular object by parametrizing its low-level feature contrast with respect to the rest of distractors and/or background. Therefore, the evaluation of saliency can be done in these terms, by accounting for feature contrast it is possible to analyze the importance to the objects that are easier to detect or that can be detected preattetively. Previous saliency benchmarks usually evaluate eye tracking data spatially across all fixations, we also propose the evaluation of saliency across fixations, which is an issue of further study. Future steps for this study would include the evaluation of saliency in dynamic scenes [175, 258] using synthetic videos with both static or dynamic camera. This would allow us to investigate the impact of temporally-variant features (e.g. motion) over saliency predictions. Another analysis to consider is the impact of the spatial location of salient features (in eccentricity terms towards the image center), which might affect each model distinctively. Each of the steps in saliency modelization (i.e. feature extraction, conspicuity computation and feature fusion) might have a distinct influence over eye movement predictions. Acknowledging that conspicuity computations are the key factor for computing saliency, a future evaluation of how each mechanism contributes to model performance might be of interest.

3.8 Conclusion

Contrary to the current state-of-the-art, we reveal that saliency models are far away from acquiring HVS performance in terms of predicting bottom-up attention. We prove this with a novel dataset SID4VAM, which contains uniquely synthetic images, generated with specific low-level feature contrasts. In this study, we show that overall Spectral/Fourier-based saliency models (i.e. HFT and WMAP) clearly outperform other saliency models when detecting a salient region with a particular conspicuous object. Other models such as AWS, CASD, GBVS and SAM-ResNet are the best predictor candidates for each saliency model inspiration categories respectively (Cognitive/Biological, Information-Theoretic, Probabilistic and Deep Learning). In particular, visual features learned with deep learning models might not be suitable for efficiently predicting saliency. Here we pose that saliency detection might not be directly related to object detection, therefore training upon high-level object features might not be significatively favorable for predicting saliency in these terms. Future saliency modelization and evaluation should account for low-level feature distinctiveness in order to accurately model bottom-up attention. Here we remark the need for analyzing other factors such as the order of fixations, the influences of the task and the psychometric parameters of the salient regions.

Visual Saliency in V1: Part II
Bottom-up attention



Extracting V1 maps of visually-conspicuous regions in still images.

4 NSWAM: Neurodynamic Saliency WAvelet Model

Initial hypotheses by Li [183, 383] suggested that visual saliency is processed by the lateral interactions of V1 cells. Here, pyramidal cells and interneurons in the primary visual cortex (V1, Brodmann Area 17 or striate cortex) and their horizontal intracortical connections modulate activity in V1. Li's neurodynamic model [184] of excitatory and inhibitory firing-rate neurons was able determine how contextual influences of visual scenes contribute to the formation of saliency. Here, interactions between neurons tuned to specific orientation sensitivities served as predictors of pop-out effects and search asymmetries [185]. Li's neurodynamic model was later extended by Penacchio et al. [239] proposing the aforementioned lateral interactions to also be responsible for brightness induction mechanisms. By considering neuron orientation selectivity at distinct spatial scales, this model can act as a contrast enhancement mechanism of a particular visual area depending of induced activity from surrounding regions.

The model is extended from previous implementation by Pennacchio et al. [239] in Matlab and C++ [1]. Here we describe the main steps in relation to the computations done to the images: 4.1.1. Feature Extraction, 4.2. Feature Conspicuity and 4.3. Feature Integration. In this section, computations in the early visual pathways will be represented in line with a stimulus example. Overall model architecture was inspired by previous work from Murray et al.'s Saliency Induction Model (SIM), also named Saliency Induction Model (SIM) [209], defining a biologically-inspired and unsupervised low-level model for saliency prediction. Although it provided a promising approach for predicting saliency maps, we aim to highlight novel computations of firing rate dynamics in accordance with physiological properties of V1 cells.

Reproducing other effects

Here we present a novel neurodynamic model of visual attention and we remark its biological plausability as being able to simultaneously reproduce other effects such as Brightness Induction [239], Chromatic Induction [65] and Visual Discom-

[1] Code can be downloaded from https://github.com/dberga/NSWAM

fort [241] effects. Brightness and Chromatic induction stand for the variation of perceived lightness and color of a visual target depending on luminance and/or chromatic properties of its surrounding area respectively. Thus, a visual target can be perceived as being different (contrast) or similar (assimilation) to its physical properties by varying its surrounding context. Visual scenes are projected to the retinal photoreceptors (RP), processed by retinal ganglion cells (RGC), and later projected from lateral geniculate nucleus (LGN) pathways towards V1 receptive fields (RF). From that, the output of V1's neuronal activity (coded as firing-rates), after several cycles of excitatory-inhibitory V1 interneuron interactions, is used as predictors of induction and saliency respectively. These responses will act as a contrast enhancement mechanism, which for the case of saliency, are integrated towards projections in the superior colliculus (SC) for eye movement control. Therewith, our model has also been able to reproducte visual discomfort, as relative contrast energy of particular region on a scene is found to produce hyperexcitability in V1 [174, 240], one of possible causes of producing certain conditions such as malaise, nausea or even migraine. Previous neurodynamic [66, 67, 82, 115, 193] and saliency models [41, 379] are able to reproduce attention processes and predict eye movements [53] but are uniquely presented to work for that specific task. On behalf of model biological plasusibility on V1 function and its computations, we present a unified model of V1 able to predict attention from real and synthetic color images while mimicking physiological properties of the neural circuitry stated previously.

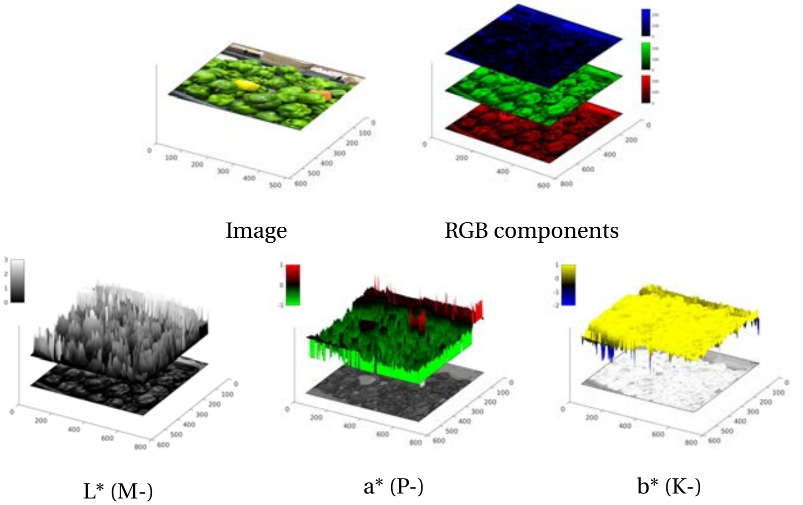
4.1 Feature Extraction

4.1.1 From Images to Sensory Signals: Retinal computations

The HVS perceives to light at distinct wavelengths of the visual spectrum and separates them to distinct channels for further processing in the cortex. First, RP (corresponding to rod and cone cells) are photosensitive to luminance (rhodopsin-pigmented) and color (photopsin-pigmented) [141, 296]. Mammal cone cells are photosensitive to distinct wavelengths between a range of $\sim$ 400–700nm, corresponding to three cell types, measured to be maximally responsive to Long (L, $\lambda_{max} \simeq 430nm$), Medium (M, $\lambda_{max} \simeq 530nm$) and Short (S, $\lambda_{max} \simeq 560nm$) wavelengths respectively [302]. RP signals are received by RGC midget, bistratified and parasol cells forming an opponent process [294] ("Red vs Green", "Blue vs Yellow", and "Light vs Dark" respectively). In order to simulate these chromatic and light intensity opponencies using digital images, we transformed the RGB color space to the CIELAB (Lab or $L^* a^* b^*$) space (including a gamma correction of γ_{RGB}=1/2.2). L^*, a^* and b^* channels represent [181] a cubic color space combining RGB value opponencies ($+L$=lighter, $-L$=darker, $+a$=reddish, $-a$=greenish, $+b$=yellowish and $-b$=blueish) as exemplified in Figure 4.1.

$$\begin{aligned} L^* &= R+G+B, \\ a^* &= \frac{R-G}{L^*}, \\ b^* &= \frac{R+G-2B}{L^*}. \end{aligned} \tag{4.1}$$

All RGB pixel values of processed images are previously corrected with $\gamma = 1/2.2$.



Image

RGB components

L* (M-)

a* (P-)

b* (K-)

Figure 4.1 – Example of CIELAB components of color opponencies given a sample image, corresponding to L^* (Intensity), a^* (Red-Green) and b^* (Blue-Yellow).

Later, receptive fields in RGC [294] are activated in a center-surround fashion, receiving ON-OFF responses, being connected to horizontal (H-cell) and bipolar cell (B-cell) upstream circuitry. B-cells are hyperpolarized (OFF) or depolarized (ON) according to RP activity. In conjunction, H-cells send excitatory (center) and inhibitory feedback (surround) to RP. Midget (R-G), bistratified (B-Y) and parasol (L-D) RGC signals are sent through the optic nerve towards Parvo- (P-), Konio- (K-) and Magno-cellular (M-) pathways in LGN.

4.1.2 Hypercolumnar organization in the brain

RGC center-surround responses are sent to LGN and projected to V1 cells. V1's cortical hypercolumns encode similar features of orientation-selective cells at different spatial frequencies. Simple cells found in V1 RFs (layers 4 & 6) are sensitive to center-surround responses at distinct orientations, whereas complex cells (found in layers 2/3, 5 & 6) overlap ON and OFF regions, and can be modeled as a combination of simple cell responses. Parvo- (P- or β), Konio- (K- or γ) and Magno-cellular (M- or α) pathways send signals separately towards distinct layers of the striate cortex (correspondingly projecting to $4C\beta$ & 6 from "P-", 2/3 & 4A from "K-" and $4C\alpha$ & 6 from "M-" cell pathways) for parallel and recurrent processing in V1.

V1 cell sensitivities to distinct orientations [137] and spatial frequencies [190] are usually modeled as Gabor filters. Since Gabor transforms cannot be inverted to obtain the original image, we used the *à trous* algorithm, which is an undecimated discrete wavelet transform (DWT) [109][301, Chapter 6]. This decomposition allows to perform an inverse, where the basis functions remain similar to Gabor filters. We propose biologically plausible computations for extracting multiple orientations and multiscale feature representations of from V1's receptive field (RF) hypercolumnar organization (Fig. 4.2). The wavelet approximation planes $c_{s,\theta}$ (s for scale and θ for orientation) are computed by convolving the image with the filter h_s.

$$\begin{aligned} c_{s,h} &= c_{s-1} \otimes h_s, \\ c_{s,v} &= c_{s-1} \otimes h'_s. \end{aligned} \tag{4.2}$$

The filter h_s is obtained from h_{s-1} by doubling its size, i.e. $h_s = \uparrow h_{s-1}$, where $\uparrow$ means upsampling by introducing zeros between the coefficients. The filter (h_s) for the first scale is

$$h_1 = \frac{1}{16}\begin{bmatrix} 1 & 4 & 6 & 4 & 1 \end{bmatrix}$$

This filter can be also transposed (h'_s) to obtain distinct approximation orientation planes $c_{s,h}$ and $c_{s,v}$. From these approximation planes, we can obtain the wavelet coefficients $\omega_{s,\theta}$ at distinct scales and orientations:

$$\begin{aligned} \omega_{s,h} &= c_{s-1} - c_{s,h}, \\ \omega_{s,v} &= c_{s-1} - c_{s,v}, \\ \omega_{s,d} &= c_{s-1} - (c_{s,h} \otimes h'_s + \omega_{s,h} + \omega_{s,v}), \\ c_s &= c_{s-1} - (\omega_{s,h} + \omega_{s,v} + \omega_{s,d}). \end{aligned} \tag{4.3}$$

Here, ω_h, ω_v and ω_d correspond to the coefficients with “horizontal", “vertical" and “diagonal" orientations. Initial $c_0 = I_o$ is obtained from the CIE L*a*b* components ($o = L^*, a^*, b^*$) and c_n corresponds to the residual plane of the last wavelet component (e.g. $s = n$). The inverse transform is obtained by integrating wavelet coefficients $\omega_{s,\theta}$ and residual planes c_n:

$$I'_o = \sum_{s=1,\theta=h,v,d}^{n} \omega_{s,\theta} + c_n. \tag{4.4}$$

Considering that for every image, $M \times N$ is the size of the feature map (resized to $N \leq 128$), wavelet coefficient scales are defined to model the spatial frequency sensitivities ($s = 1..S$), where $S = \lfloor log_2(N/8) \rfloor + 2$. From these equations, three orientation selectivities can be extracted, corresponding to horizontal ($\theta_h \simeq \{0 \pm 30 || 180 \pm 30\}^o$), vertical ($\theta_v \simeq \{90 \pm 30 || 270 \pm 30\}^o$) and diagonal ($\theta_d \simeq \{45 \pm 15 || 135 \pm 15 || 225 \pm 15 || 315 \pm 15\}^o$) angles. For the case of scale features, sensititivies to size (in degree of visual angle) correspond to $2^{s_0(s-1)}/\{pxva\}$, where "$pxva$" is the number of pixels for each degree of visual angle according to experimentation (approximately between 35 and 40 px), and s_0=8, is the minimum size of the wavelet filter (h_0) defining the first the scale frequency sensitivity.

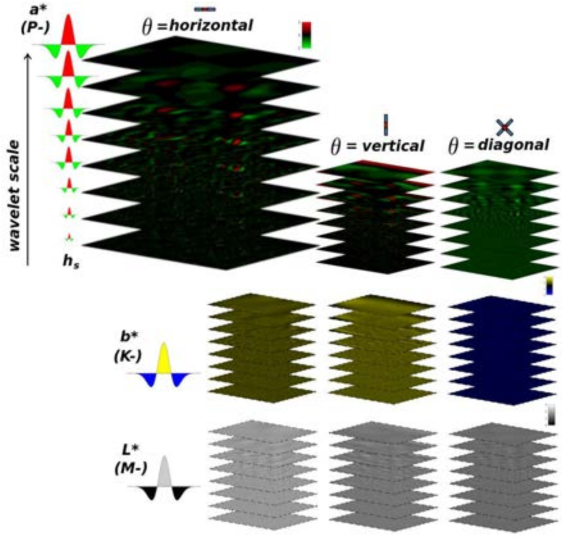


Figure 4.2 – Representation of wavelet coefficients ($\omega_{iso\theta}$), in conjunction with the output of "a-trous" wavelet transform applied to components ($o = L^*, a^*, b^*$) shown in Figure 4.1.

4.2 Feature Conspicuity

4.2.1 Computation of lateral Interactions in V1 cells

Li's hypotheses suggest that V1 computations are responsible of generating a bottom-up saliency map[183, 383]. These hypotheses state that intracortical interactions between orientation-selective neurons in V1 are able to explain contextually-dependent perceptual effects present in pre-attentive vision [184, 185, 186, 382, 385, 386], relative to contour integration, visual segmentation, visual search asymmetries, figure-ground and border effects, among others. Pop-out effects that form the saliency map are the result of horizontal connections in V1, that interact with each other locally and reciprocally. These connections are formed by excitatory cells and inhibitory interneurons [105, 347], processing information from pyramidal cell signals in layers of V1. Spatial organization of these cells accounts for selectivity in their orientation columns, their RF size and axonal field localization. The aforementioned interactions between orientation-selective cells was defined by Li's model [184] of excitatory-inhibitory firing-rate neural dynamics, later extended by Penacchio et al. [239]. Here, contrast enhancement or suppression in neural responses emerge from lateral connections as an induction mechanism. Latest implementation done by Berga & Otazu [26] for saliency prediction used colour images, where chromatic (P-,K-) and luminance (M-) opponent channels were individually processed in order to compute firing-rate dynamics of each pathway separately. With cortical magnification, each gaze can significantly vary contextual information and therefore the output of the model (we started the first view at the center of the image). Our excitatory-inhibitory model is described in Table 5.1. Horizontal connections (lateral and reciprocal) are schematized in Figure 4.3 and Table 5.1C, where excitatory cells have self-directed (J_0) and monosynaptic connections (J) between each other, whereas dysynaptically connected through (W) inhibitory interneurons. Axonal field projections follow a concentric toroid of radius $\Delta_s = 15 \times 2^{s-1}$ and radial distance Δ_θ (accounting for RF size d_s and radial distance β). Membrane potentials of excitatory ($\dot{x}_{is\theta}$) and inhibitory ($\dot{y}_{is\theta}$) cells are obtained with partial derivatives defined in Table 5.1D, composed by a chain of functions that consider firing-rates (obtained by piece-wise linear functions g_x and g_y) and membrane potentials from previous membrane cycles (modulated by α_x, α_y constants), current lateral connection potentials (J and W) and spread of inhibitory activity within hypercolumns (ψ). Background inputs (I_{noise} and I_{norm}) correspond to a simulation of random noise and divisive normalization signals (i.e. accounting for local nonorientation-specific cortical normalization and nonlinearities). Further details of model equations and parameters are specified in Table 5.1 and [239, Supporting Information S1].

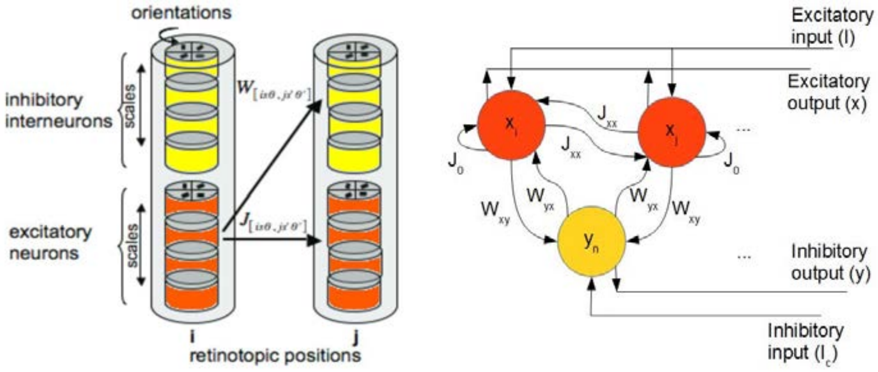


Figure 4.3 – **Left:** Representation of cortical hypercolumns with scale and orientation selectivity interactions. **Right:** Model's intracortical excitatory-inhibitory interactions, membrane potentials (orange "$\dot{x}$" for excitatory and yellow "$\dot{y}$" for inhibitory) and connectivities ("J" for monosynaptic excitation and "W" for dysynaptic inhibition).

Input signals ($I^t_{i;so\theta}$) have been defined as the wavelet coefficients ($\omega^t_{iso\theta}$), splitted between ON and OFF components (representing ON and OFF-center cell signals from RGC and LGN) depending of the value polarity (+ for positive and − for negative coefficient values) from the RF. These signals are processed separately during 10τ ($\tau = 1$ membrane time $= 10ms$; $\tau \equiv 10$ cycles), including a rest interval (empty input) of 3τ to simulate intervals between each saccade shift. The model output has been computed as the firing-rate average g_x of the ON and OFF components ($M(\omega^{t+}_{iso\theta})$ and $M(\omega^{t-}_{iso\theta})$) during the whole viewing time, corresponding to a total of 10 membrane time (being the mean of g_x for a specific range of t). By combining the outputs of all components (Equation 4.5), we can describe the changes of the model (resulting from the simulated lateral interactions of V1) with respect the original wavelet coefficients $\omega^t_{iso\theta}$, which alternatively defines the contrast enhancement seen on the brightness and chromatic induction cases. Our result ($S^t_{i;o}$) will define the saliency map as an average conspicuity map or feature-wise distinctiveness (RF firing rates across scales and orientations for each pathway). In our previous work [65, 239, 241], the final model output is obtained by combining the model result with the wavelet coefficients ($M(\omega^t_{iso})\omega^t_{iso}$) instead. Considering that for every image, $M \times N$ is the size of the feature map (resized to $N \leq 128$), wavelet coefficient scales are defined to model the spatial frequency sensitivities ($s = 1..S$), where $S = \lfloor log_2(N/8) \rfloor + 2$. Top-down selection can be introduced to the model as an inhibitory control (I_c) mechanism, further explained in Table 5.1E and in the Section 5.3.

$$\hat{S}^t_{i;o} = \sum_{s=1..S;\theta=h,v,d}^{n_s} M(\omega^{t+}_{iso\theta}) + \sum_{s=1..S;\theta=h,v,d}^{n_s} M(\omega^{t-}_{iso\theta}) + c_i, \tag{4.5}$$

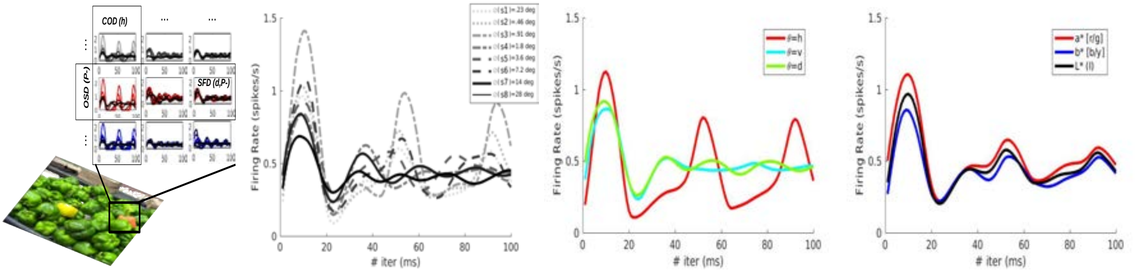


Figure 4.4 – Firing rates plotted for 10 membrane time (100 iterations) accounting for neurons (ON+OFF values) inside a specific region (1st col.). Mean firing rates for all scales (Spatial Frequency Dynamics, 2nd col.), orientations (Orientation Selectivity Dynamics, 3rd col.), and color channels (Chromatic Opponency Dynamics, 4th col.).

4.3 Feature Fusion/Integration

4.3.1 Generating the saliency map in the Superior Colliculus

Latest hypotheses about neural correlates of saliency [336, 352] state that the superior colliculus is responsible of encoding visual saliency and to guide eye movements [281, 350]. Acknowledging that the superficial layers of the SC (sSC) receive inputs from the early stages of visual processing (V1, retina), the SC selects these as the root of bottom-up activity to be selected in the intermediate and deep layers (iSC, dSC). In accordance to the previous stated hypotheses[183, 383], saccadic eye movements modulated by saliency therefore are computed by V1 activity, whereas recurrent and top-down attention is processed by neural correlates in the parieto-frontal cortex and basal ganglia. All these projections are selected as a winner-take-all mechanism in SC[183, 185] to a unique map, where retinotopic positions with the highest activity will be considered as candidates to the corresponding saccade locations. These activations in the SC are transmitted to guide vertical and horizontal saccade visuomotor nerves [128].

The behavioral quantity of the unique 2D saliency map has been defined by computing the inverse of the previous processes using the model output for each pathway separately. Retinotopic positions have been transformed to coordinates in the visual space using the inverse of the cortical magnification function (Equation 5.2). Output signals (V1 sensitivities to orientation and spatial frequencies) are integrated by computing the inverse discrete wavelet transform to obtain unique maps for each channel opponency (Equation 4.4). A unique representation (Equation 4.6) of final neuronal responses for each pathway (P-, K- and M- as a^*, b^* and L^*) is generated with the euclidean norm as in Murray et al.[208, 209] model. The resulting map is later normalized by the variance (Equation 4.7) of the firing rate [383, Chapter 5]. This map represents the final saliency map, that describes the probability distribution of fixation points in certain areas of the image. In addition to this estimation, the saliency map has been convolved with a gaussian filter simulating a smoothing caused by the deviations of $\sigma = 1$ deg given from eye tracking experimentation, recommended by LeMeur & Baccino [180].

$$\hat{S}_i = \sqrt{\hat{S}_{i;a^*} + \hat{S}_{i;b^*} + \hat{S}_{i;L^*}}, \tag{4.6}$$

$$z_i(\hat{S}) = \frac{\hat{S}_i - \mu_{\hat{S}}}{\sigma_{\hat{S}}}, \tag{4.7}$$

4.4 Evaluation Metrics for Saliency

Fixations and saccades are captured using eye-tracking technology. Eye movement data is combined across all fixations from participants' data, being represented as binary maps (called fixation maps), according to the fixation localizations in the visual space for each corresponding image, or as density distributions (alternatively named density maps) from these fixations considering eye-movement localization probabilities (Figure 4.6). Fixation density maps are computed accordingly from fixation maps with a gaussian filter [180].

Prediction scores are calculated using spatially-dependent metrics [55][54] which compare either fixation maps or fixation density maps to saliency map predictions from the models (AUC, CC, NSS, KL and SIM). Essentially, these metrics assign a score considering true positive (TP) values for the saliency predictions inside the locations from the fixation maps (or higher correlations with respect to density maps) and false positive (FP) values for the reverse cases. Other metrics compare saliency maps with a baseline set of other image fixation maps in order to prevent behavioral tendencies such as center biases, which are not representative data for saliency prediction. Similarly, a baseline gaussian of all images is used (InfoGain) for minimizing center biases on prediction scores.

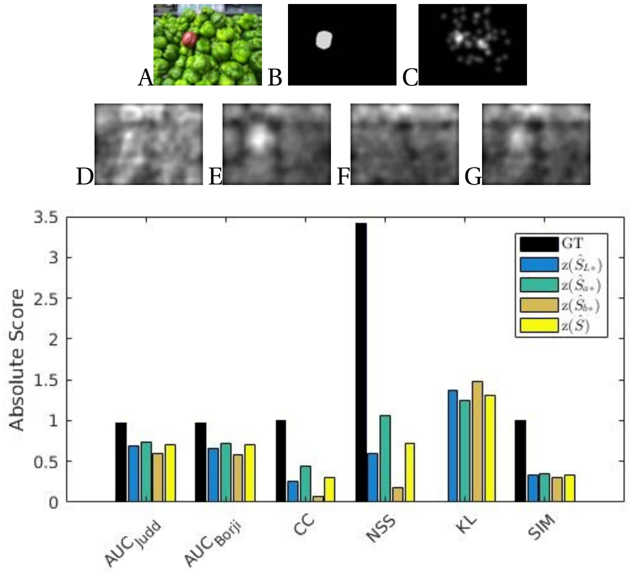


Figure 4.5 – **(A)** Example Image. **(B)** Mask of salient region. **(C)** Fixation density map (GT). **(D,E,F,G)** Predicted saliency map given $z(\hat{S}_{i;L^*})$, $z(\hat{S}_{i;a^*})$, $z(\hat{S}_{i;b^*})$ and $z(\hat{S}_i)$ respectively. **(E)** Results of prediction metrics from these saliency maps. $z(\hat{S})$ corresponds to our model's saliency prediction (NSWAM).

4.5 Results

4.5.1 Predicting human eye movements

We have computed the saliency maps[2] for images from distinct eye-tracking datasets, corresponding to 120 real scenes (Toronto) [50], 40 nature scenes (KTH) [160], 100 synthetic patterns (CAT2000$_{Pattern}$)[42] and 230 synthetic images with specific feature contrast (SID4VAM) [23]. We have computed these image datasets with deep supervised artificial saliency models that specifically compute high-level features (OpenSalicon [135][313], DeepGazeII [171], SAM [76], SalGan [230]), and models that extract low-level features, corresponding to the cases with artificial (SUN [380], GBVS [119]) and biological inspiration (IKN [142], AIM [49], SSR [286], AWS [102] and SIM [209]). The Saliency WAvelet Model (SWAM) and Neurodynamic SWAM

[2] Code for model evaluations can be downloaded from https://github.com/dberga/saliency

(NSWAM) corresponds to our model excluding or including lateral interactions shown in 4.2.

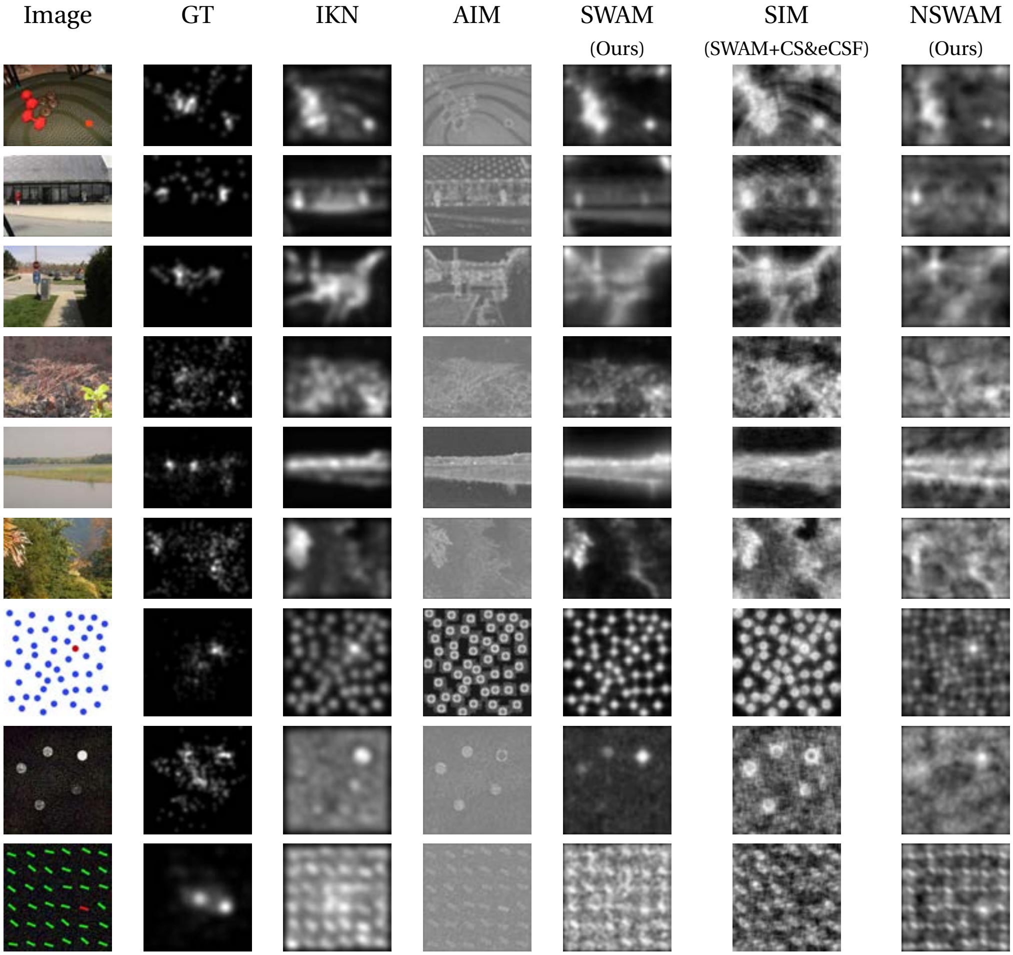


Figure 4.6 – Examples of saliency maps from Itti et al. (IKN), Bruce & Tsotsos (AIM), Saliency WAvelet Model (SWAM), Murray et al.'s model (SIM) and our Neurodynamic model (columns 3 to 7, respectively), corresponding to images with distinct contexts (column 1). We also show the density distribution of fixations given by the eye-tracking experimentation (column 2).

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.969**	**0.954**	**1.000**	**3.831**	**0.000**	**1.000**	**0.903**	**2.425**
HIGH-LEVEL	**OpenSalicon**	0.821	0.771	0.522	1.655	1.113	0.429	0.716	0.232
	DeepGazeII	0.850	0.768	0.595	1.877	0.997	0.483	**0.717**	**0.422**
	SAM$_{RESNET}$	0.850	0.725	0.612	1.955	2.420	**0.516**	0.666	-1.555
	SAM$_{VGG}$	0.569	0.543	0.055	0.158	11.972	0.214	0.506	-15.522
	SalGan	**0.858**	**0.816**	**0.629**	**1.898**	**0.986**	0.510	0.716	0.387
LOW-LEVEL	**SUN**	0.694	0.682	0.242	0.755	1.589	0.290	0.645	-0.499
	GBVS	0.817	0.803	0.487	1.431	1.168	0.397	0.632	0.077
	SSR	0.765	0.756	0.364	1.084	1.355	0.340	0.700	-0.174
	AWS	0.773	0.761	0.401	1.229	1.322	0.352	0.714	-0.106
	AIM	0.727	0.716	0.292	0.883	1.612	0.314	0.663	-0.580
	IKN	0.794	0.782	0.421	1.246	1.248	0.366	0.650	-0.024
	SIM	0.754	0.744	0.317	0.951	1.486	0.302	0.705	-0.369
	SWAM (Ours)	***0.728***	***0.716***	***0.287***	***0.868***	***1.492***	***0.305***	***0.654***	***-0.378***
	NSWAM (Ours)	***0.706***	***0.694***	***0.257***	***0.764***	***1.604***	***0.278***	***0.631***	***-0.552***

Table 4.1 – Results for prediction metrics with Toronto dataset [49], corresponding to real (indoor and outdoor) images.

	method	↑AUC$_{Judd}$	↑AUC$_{Borji}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.902**	**0.850**	**1.000**	**2.038**	**0.000**	**1.000**	**0.822**	**1.415**
HIGH-LEVEL	**OpenSalicon**	0.634	0.611	0.300	0.452	0.780	0.541	0.556	-0.278
	DeepGazeII	0.648	0.618	0.362	**0.578**	**0.678**	0.559	**0.588**	**-0.104**
	SAM$_{RESNET}$	**0.660**	0.599	0.371	0.570	3.125	0.508	0.548	-3.643
	SAM$_{VGG}$	0.525	0.525	0.058	0.074	8.800	0.354	0.501	-11.836
	SalGan	0.655	0.626	**0.391**	0.581	1.666	0.544	0.560	-1.554
LOW-LEVEL	**SUN**	0.535	0.532	0.083	0.132	0.804	0.512	0.526	-0.303
	GBVS	0.649	**0.638**	0.351	0.505	0.711	0.563	0.533	-0.177
	SSR	0.575	0.573	0.172	0.270	0.778	0.525	0.557	-0.260
	AWS	0.587	0.583	0.210	0.329	0.851	0.511	0.581	-0.362
	AIM	0.572	0.568	0.179	0.274	0.918	0.523	0.552	-0.509
	IKN	0.617	0.611	0.274	0.403	0.714	0.547	0.551	-0.173
	SIM	0.587	0.584	0.201	0.311	0.745	0.531	0.573	-0.212
	SWAM (Ours)	***0.601***	***0.596***	***0.231***	***0.346***	***0.749***	***0.529***	***0.574***	***-0.221***
	NSWAM (Ours)	***0.598***	***0.593***	***0.230***	***0.345***	***0.711***	***0.536***	***0.565***	***-0.168***

Table 4.2 – Results for prediction metrics with KTH dataset [160] subset of uniquely nature images.

Our results show that our model has a trend to acquire other saliency models performance, with an emphasis on outperforming previous Murray's SIM model for the cases of KTH, CAT2000 and SID4VAM (Tables 4.2, 4.3 and 4.4), corresponding to nature and synthetic images, as well as showing stable metric scores for distinct contexts (similarly as AWS and GBVS). NSWAM outperforms other biologically-inspired models (IKN, AIM, SSR, SWAM & SIM) specially for metrics that account for center biases. These center biases are qualitatively present even for images where the

	method	↑AUC_{Judd}	↑$\text{AUC}_{\text{Borji}}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.895**	**0.826**	**0.890**	**2.335**	**0.265**	**0.736**	**0.623**	**0.777**
HIGH-LEVEL	**OpenSalicon**	0.651	0.621	0.220	0.603	1.526	0.357	0.555	-1.092
	DeepGazeII	0.611	0.561	0.157	0.467	1.932	0.325	0.547	-1.657
	$\text{SAM}_{\text{RESNET}}$	**0.766**	0.711	**0.518**	**1.356**	1.747	**0.456**	0.546	-1.444
	SAM_{VGG}	0.625	0.581	0.123	0.320	8.581	0.322	0.508	-11.262
	SalGan	0.751	0.714	0.417	1.080	1.720	0.430	0.553	-1.384
LOW-LEVEL	**SUN**	0.549	0.539	0.068	0.193	5.860	0.280	0.526	-7.237
	GBVS	0.759	**0.717**	0.399	1.056	**1.113**	0.430	0.561	**-0.503**
	SSR	0.592	0.582	0.118	0.318	1.760	0.334	0.568	-1.432
	AWS	0.604	0.594	0.209	0.609	1.521	0.339	**0.595**	-1.077
	AIM	0.570	0.565	0.118	0.332	5.323	0.301	0.544	-6.490
	IKN	0.701	0.692	0.323	0.828	1.267	0.382	0.562	-0.724
	SIM	0.586	0.578	0.120	0.336	1.614	0.328	0.566	-1.225
	SWAM (Ours)	***0.617***	***0.602***	***0.180***	***0.503***	***1.484***	***0.335***	***0.571***	***-1.029***
	NSWAM (Ours)	**0.588**	***0.584***	***0.139***	***0.383***	***1.471***	***0.326***	***0.571***	***-1.017***

Table 4.3 – Results for prediction metrics with CAT2000 dataset [42] training subset (Pattern) of uniquely synthetic images.

	method	↑AUC_{Judd}	↑$\text{AUC}_{\text{Borji}}$	↑CC	↑NSS	↓KL	↑SIM	↑sAUC	↑InfoGain
	Humans	**0.943**	**0.882**	**1.000**	**4.204**	**0.000**	**1.000**	**0.860**	**2.802**
HIGH-LEVEL	**OpenSalicon**	0.692	0.673	0.284	0.956	1.549	0.375	0.615	0.052
	DeepGazeII	0.640	0.634	0.177	0.630	1.685	0.336	0.618	-0.150
	$\text{SAM}_{\text{RESNET}}$	0.727	0.673	0.305	0.967	2.610	0.388	0.600	-1.475
	SAM_{VGG}	0.537	0.523	0.026	0.070	11.947	0.216	0.503	-14.954
	SalGan	0.715	0.662	0.287	0.883	2.506	0.373	0.593	-1.350
LOW-LEVEL	**SUN**	0.542	0.532	0.080	0.333	16.408	0.165	0.530	-21.024
	GBVS	**0.747**	**0.718**	**0.400**	**1.464**	**1.363**	**0.413**	0.628	**0.331**
	SSR	0.672	0.665	0.192	0.639	1.904	0.365	0.642	-0.467
	AWS	0.679	0.667	0.255	1.088	1.592	0.373	**0.672**	0.013
	AIM	0.570	0.566	0.122	0.473	14.472	0.224	0.557	-18.182
	IKN	0.686	0.678	0.283	0.878	1.748	0.380	0.608	-0.233
	SIM	0.650	0.641	0.189	0.694	1.702	0.357	0.619	-0.148
	SWAM (Ours)	***0.639***	***0.618***	***0.177***	***0.682***	***1.799***	***0.340***	***0.601***	***-0.281***
	NSWAM (Ours)	***0.614***	***0.610***	***0.136***	***0.529***	***1.686***	***0.335***	***0.622***	***-0.150***

Table 4.4 – Results for prediction metrics with SID4VAM dataset [23] with synthetic images.

salient region is conspicuous Fig. 4.6, rows 8 & 9. Saliency models that compute high-level visual features are shown to perform better with real image scenes (Table 4.1). However, the image contexts that lack of high-level visual information should be more representative indicators of saliency, due to the absence of semantically or contextually-relevant visual information (nature images), or to be characterized to uniquely contain low-level features (synthetic images) presenting clear pop-out spots to direct participants fixations (which would cause lower inter-participant differences and therefore lower center biases). Although AWS and GBVS perform better on predicting fixations at distinct contexts, we remark the plausibility of our unified design for modeling distinct HVS' functionality. NSWAM shows a new insight of applying a more biologically plausible computation of the aforementioned steps. First, we transform image values to color opponencies, found in RGC. Second, we model LGN projections to V1 simple cells using a multiresolution wavelet transform. Third, conspicuity is computed with a dynamical model of the lateral interactions of V1. Fourth, these channels are integrated to a unique map which will represent SC activity. Using a neurodynamic model with firing-rate neurons allows a more detailed understanding of the dependency of saliency on lateral connections and a potential further study in terms of single neuron dynamics using real image scenes.

4.5.2 Psychophysical measurements

Acknowledging that the HVS processes information according to the context, human performance on detecting a salient object on a scene may also vary according to the visual properties of such object. With a synthetic image dataset [23] a specific analysis of how each individual feature influences saliency can be done. In this study we will show how fixation data is predicted when varying feature contrast, concretely on parametrizing Set Size, and Brightness, Color, Size and Orientation contrast between a target salient object Fig. **??**B and the rest of distractors (feature singleton search). In this section a set of psychophysical stimuli will be displayed with its parametrization of set size or feature contrast and its sAUC in comparison to other biologically-inspired saliency models.

The shuffled AUC (sAUC) is the metric we used for our psychophysical experimentation. It computes the area under ROC considering TP as fixations inside the saliency map, similarly to the AUC. However, this metric does not evaluate FP at random areas of the image but instead uses fixations inside other random images from the same dataset over several trials (10 by default). This metric gives a more accurate evaluation of predicted maps with respect human fixations but penalizing for higher model center biases (which are or can be present for distinct images in the ground truth).

Brightness differences

Differences in brighness are major factors for making an object to attract attention. Thus, a bright object is less salient as luminance of other objects increase (Fig. 4.7). Conversely, a dark target in a bright background will be more salient as distractors have higher luminance [238][218]. NSWAM processes luminance signals separately from chromatic ones using the L* channel (feature conspicuity from a distinctively bright object upon a dark background will be processed similarly to a dark object upon a bright background). We compare sAUC metrics for both conditions and NSWAM is shown to acquire similar performance to SIM and SWAM, with higher sAUC than IKN Fig. 4.8,A-B, specially for stimulus with higher contrasts ($\Delta L_{D,T} > .25$). Results on sAUC for NSWAM correlates with brightness contrast, for both cases of bright ($\rho = .941, p = 1.6 \times 10^{-3}$) and dark ($\rho = .986, p = 4.7 \times 10^{-5}$) background.

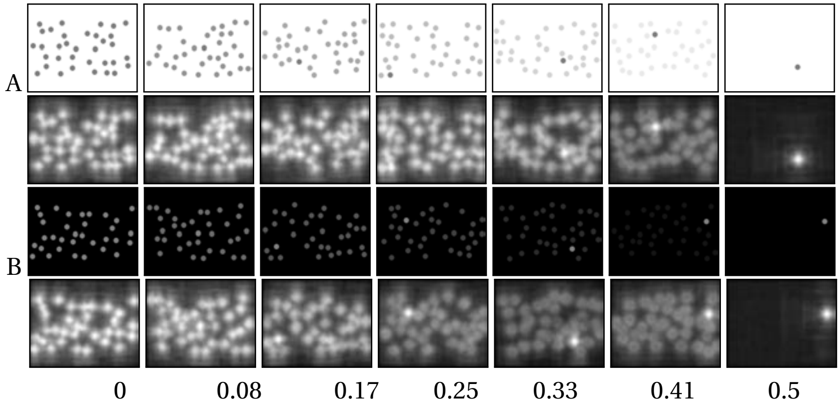
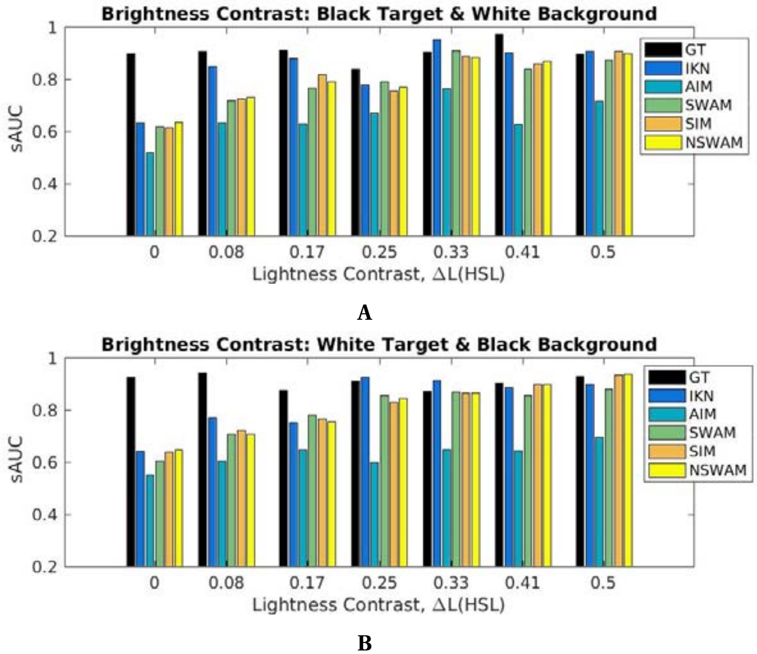


Figure 4.7 – Array of synthetic stimuli representing distinct brightness contrasts (HSL luminance differences) from target and distractors ($\Delta L_{D,T}$) with **(A)** bright background ($L_T = 0.5, L_B = 1, L_D = 0.5..1$) and conversely, with **(B)** dark background ($L_T = 0.5, L_B = 0, L_D = 0..0.5$). Rows below **A,B** correspond to NSWAM's predicted saliency maps.



A

B

Figure 4.8 – Results of sAUC upon brightness contrast on a luminance singleton ($\Delta L_{D,T}$) with a **(A)** bright and **(B)** dark background.

Color differences

Color changes spatial and temporal behavior of eye movements, influencing how conspicuous are specific objects on a scene [87][17]. Similarly to previous section, here we vary the chromaticity of the background, which can alter search efficiency [210][79]. In this section, we used stimuli similar to Rosenholtz's experimentation [268], with red and blue singletons for achromatic or oversaturated backgrounds Fig. 4.9. Here, chromatic contrast is defined as the HSL saturation differences ($\Delta S_{D,T}$) between a salient target and the rest of distractors.

Similarly to Fig. 4.7, NSWAM has similar sAUC to SIM for all background conditions (Fig. 4.10,A-D). Achromatic backgrounds contribute to salient object detection by increasing sAUC of the pop-out singleton. That effect is present for visual search results and our saliency prediction. Results comparing target search fixation maps and sAUC show distinct performance upon saturation contrast depending on background conditions. Cases which stimulus background was achromatic,

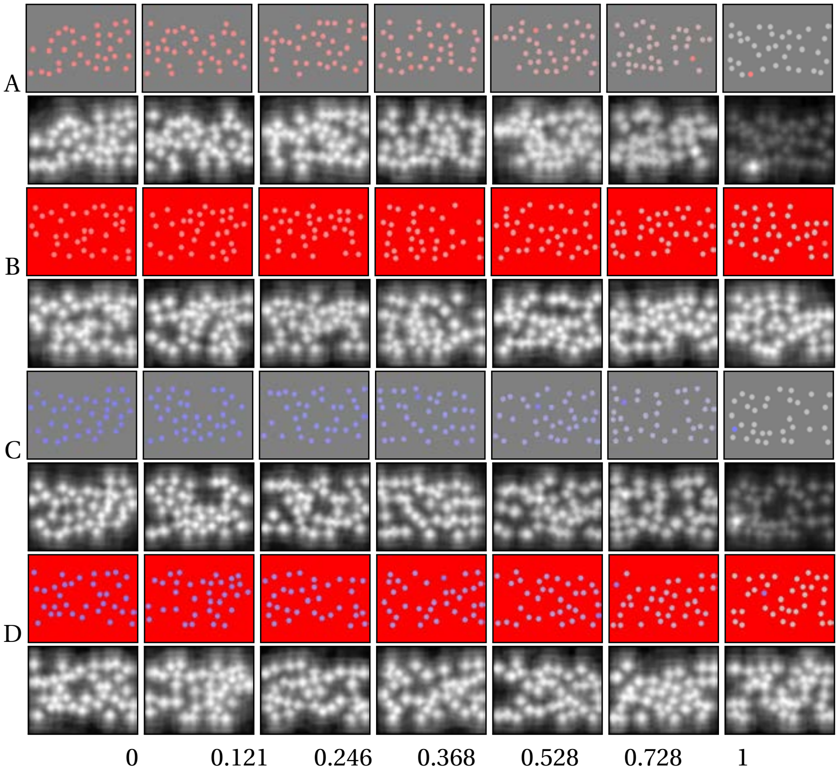
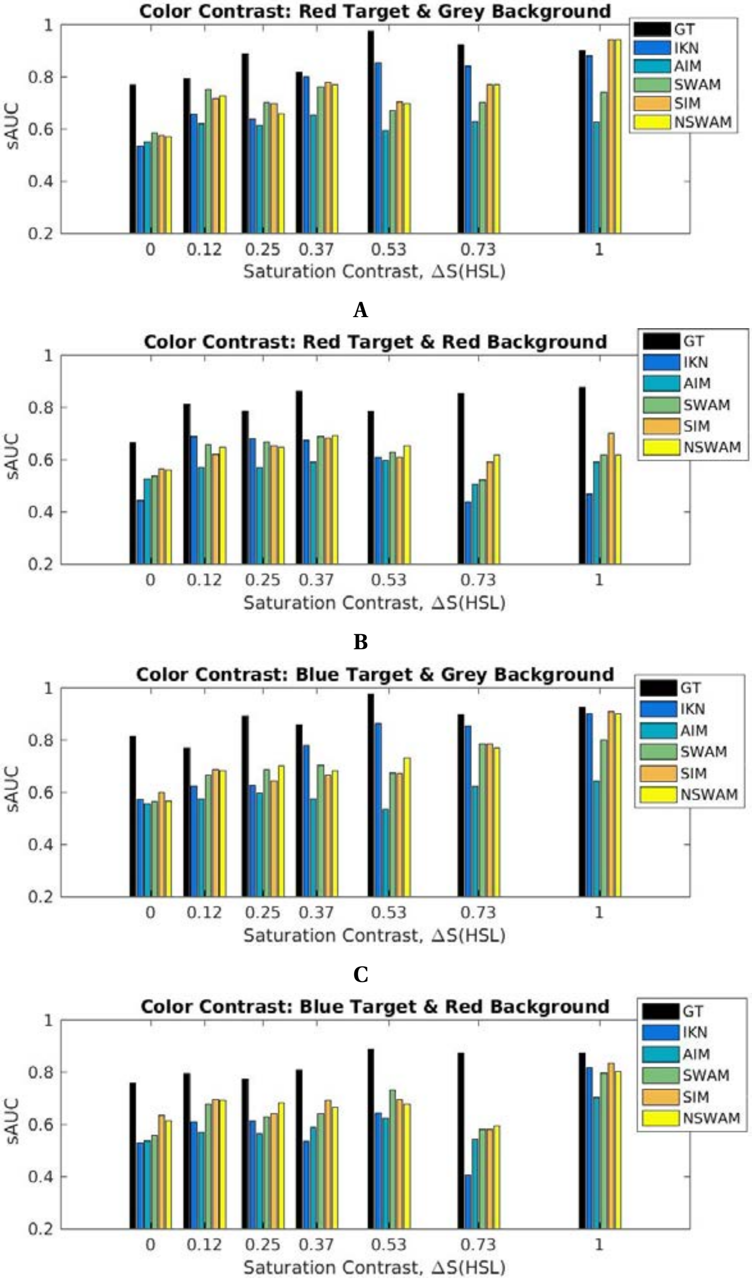


Figure 4.9 – Chromatic stimuli upon saturation contrast ($\Delta S_{D,T}$) between a red target ($H_T = 0$ž) and an **(A)** unsaturated, grey background or an **(B)** oversaturated, red background. Other cases **(C,D)** present a blue target ($H_T = 240$ž) with same background properties to **(A)** and **(B)** respectively. Rows below **A-D** correspond to NSWAM's predicted saliency maps.

distinct from the feature singleton, had higher correlation than with oversaturated background. For the cases of grey (achromatic) background, there is a correlation between sAUC results for our model and $\Delta S_{D,T}$ with a red ($\rho = .864, p = 1.2 \times 10^{-2}$) and blue ($\rho = .944, p = 1.4 \times 10^{-3}$) target singleton. However, when background color is oversaturated red and targets are either red ($\rho = .106, p = .82$) or blue ($\rho = .483, p = .27$), then saturation contrast do not correlate with sAUC.



A

B

C

D

Figure 4.10 – Results of sAUC upon saturation contrast ($\Delta S_{D,T}$) on a red singleton with **(A)** achromatic or **(B)** oversaturated red background, or either a blue singleton with **(C)** achromatic or **(D)** oversaturated red background.

Size contrast

Feature distinctiveness with feature singletons have been tested by varying set size, object orientation and/or color. Here is tested how object size affects its saliency, previously tested with visual search experimentation [112][310][248]. A set of 34 symmetric objects (with a dark circle shape) are distributed randomly around the image Fig. 4.11, preserving equal diameter. One of the circles is defined with dissimilar size, either with higher or lower diameter with respect the rest (which are defined with a diameter of 2.5 deg). Performance for NSWAM's sAUC improves with size dissimilarity. When the diameter of the dissimilar circle is higher, sAUC is higher for that particular region. For the highest scaling factor (when the dissimilar object is bigger), NSWAM has higher sAUC compared to previous biologically-inspired models (Fig. 4.12). Plus, there is a significant correlation between circle diameter and our model's results of sAUC ($\rho = .955, p = 8.3 \times 10^{-4}$).

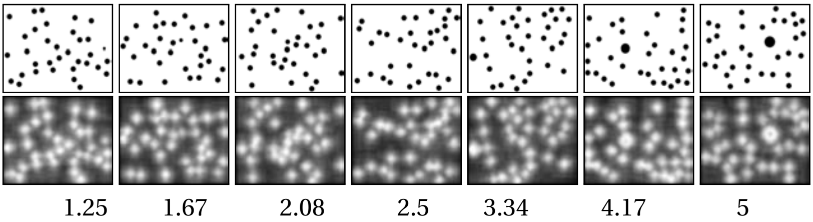


Figure 4.11 – Examples of circle distractors with equal diameter ($\varnothing_D$=2.5deg), containing a salient one with dissimilar size ($\varnothing_T$=1.25..5deg) with respect the rest. In lower row there are NSWAM's predicted saliency maps.

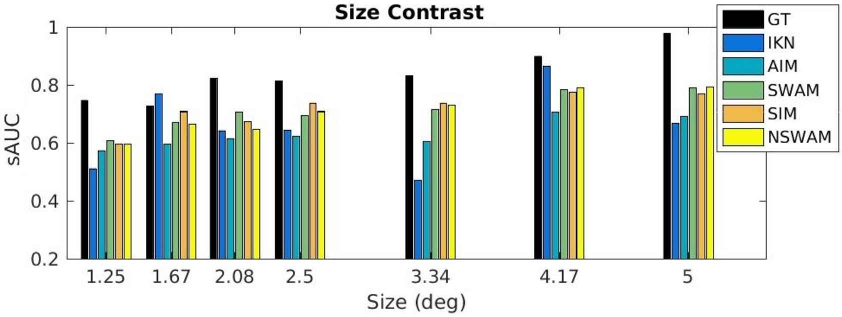


Figure 4.12 – sAUC results for Size Contrast stimuli.

Orientation contrast

Using a similar setting, varying angle of objects is found to increase search efficiency when angle contrast is increased [86][217][216]. A total of 34 bars were oriented horizontally and randomly displaced around the scene (Fig. 4.13). The dissimilar object for this case is a bar oriented with an angle contrast with respect the rest of bars of $\Delta\Phi(1,0)$=[0, 10, 20, 30, 42, 56, 90]º. Although results of sAUC show that NSWAM overperforms SIM's saliency maps, IKN is best for capturing orientation distinctiveness (Fig. 4.14). In NSWAM, 3 types of orientation selective cells are modeled, corresponding to the orientation for the wavelet coefficients ($\theta = h, v, d$). A higher number of orientation selective cells would provide a higher accuracy, specially for diagonal angles (here we only provide $\theta = d$ for 45/135º combined). By modeling orientation selective cells with 2D Gabor and Log-Gabor transforms [179][95][102] it would be possible to correctly build an hypercolumnar organization with a higher number of angle sensitivities.

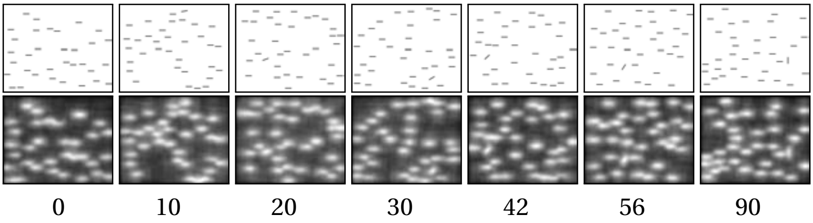


Figure 4.13 – An oriented bar with an orientation contrast of $\Delta\Phi = 0..90$ with respect to a set of bars oriented at $\Phi_D = 0$. In lower row there are NSWAM's predicted saliency maps.

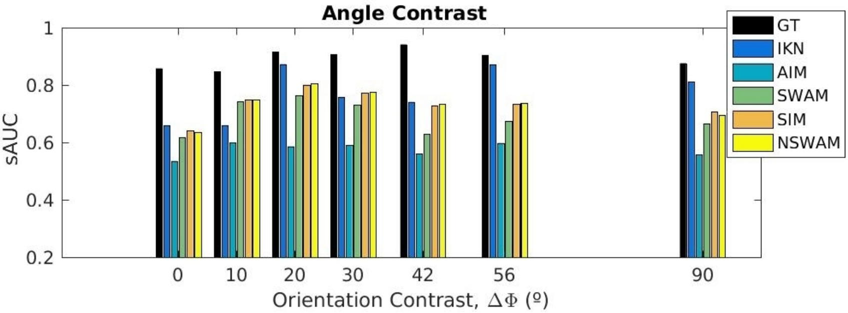


Figure 4.14 – sAUC results for Orientation Contrast stimuli.

We have to acknowledge that for this experimentation, distractors have been set with same horizontal configuration. Specific connectivity interactions [11] between orientation dissimilarities needs to be defined in order to reproduce orientation-dependent visual illusions and conspicuousness under heterogeneous, nonlinear and categorical angle configurations (seen to be done by V2 cells [10]), which are previously known to distinctively affect visual attention [217][216][101].

Visual Asymmetries

Search asymmetries appear when searching target of type “a" is found efficiently among distractors of type “b", but not in the opposite case (i.e. searching for "b" among distractors of type “a") [320][363]. Previous studies pointed out this this concept when searching a circle crossed by a vertical bar among plain circles and searching a plain circle among circles crossed by a vertical bar). Using these two configurations, we filled a grid of distractors according to specific scales (Fig. 4.15). Scale values ($s = [1.25, 1.67, 2.08, 2.5, 3.33, 4.17, 5]$ deg) change the amount of items, with arrays of 5 × 7, 6 × 8, 8 × 10, 10 × 13, 15 × 20 and 20 × 26 objects. In Fig. 4.8 our model is not only more efficient than other biologically-inspired models upon dissimilar sized objects but also on detecting conspicuous objects at distinct scales, accounting for lower or larger amount of distractors.

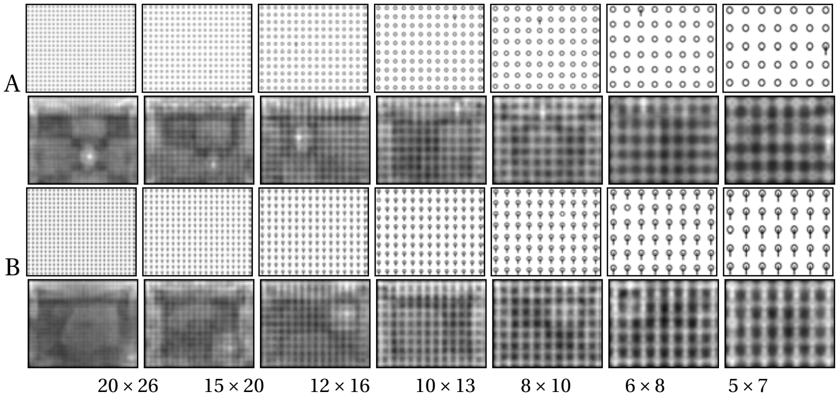
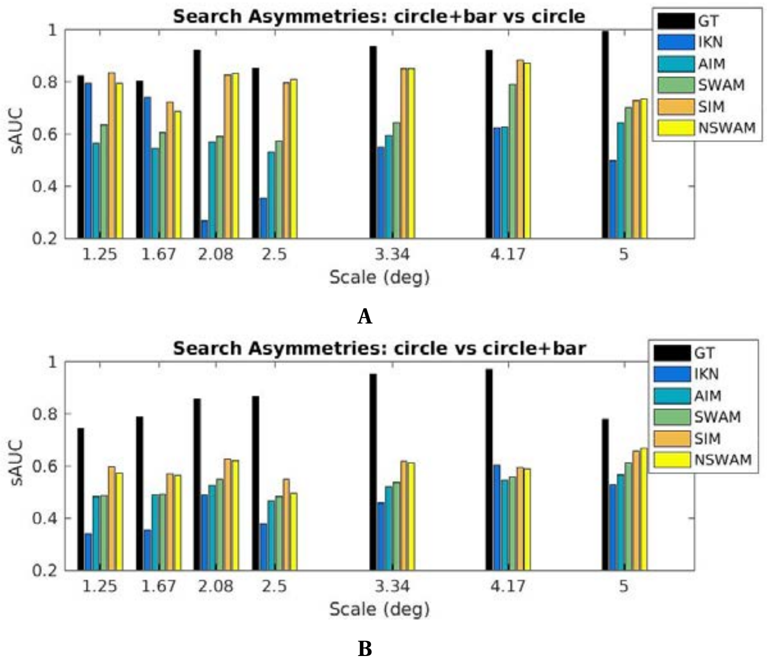


Figure 4.15 – Stimuli with distinct set sizes corresponding to search asymmetries present on a **(A)** salient circle crossed by a vertical bar among other circles and a **(B)** salient circle among other circles crossed by a vertical bar. Rows below **A,B** correspond to NSWAM’s predicted saliency maps.



Search Asymmetries: circle+bar vs circle

A

Search Asymmetries: circle vs circle+bar

B

Figure 4.16 – Results of sAUC upon varying scale and set size of **(A)** an array of circles and a salient one crossed by a vertical bar and **(B)** an array of circles crossed by a bar and a salient circle.

sAUC for NSWAM showed to correlate for a conspicuous circle crossed by a vertical bar among circles ($\rho = .83$, $p = 2.1 \times 10^{-2}$) but not for a conspicuous circle among circles crossed by a vertical bar ($\rho = .15$, $p = .75$).

4.6 Conclusion

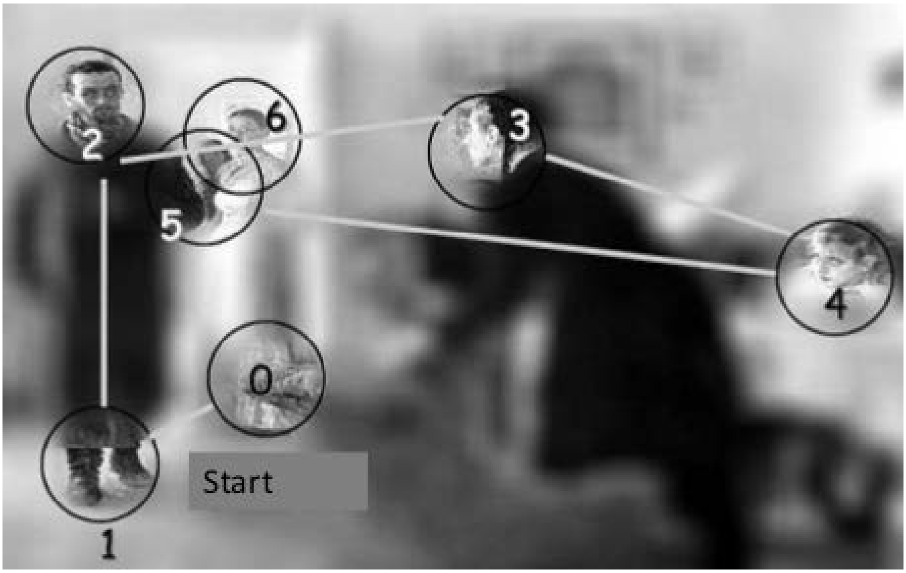
In this work, we hypothesize that low-level saliency is likely to be associated by the computations of V1. For such, we proposed a neurodynamic model of V1's lateral interactions, processing each channel separately and acquiring firing rate dynamics from real image simulations. Here we have to pinpoint three statements in agreement with our findings:

- First, our model of the lateral interactions in V1 has a trend to acquire state-of-the-art results on human eye fixations, specifically, with natural and synthetic images.

- Second, our model improves results for biologically-inspired saliency models and it is consistent with human psychophysical measurements (tested for Visual Asymmetries, Brightness, Color, Size and Orientation contrast). Adding up to the stated hypothesis, our model presents highest performance at highest contrast from feature singleton stimuli (where salient objects pop-out easily).
- Three, we remark the model plausibility by mimicking HVS physiology on its processing steps and being able to reproduce other effects such as Brightness Induction [239], Color Induction [65] and Visual Discomfort [241], efficiently working without applying any type of training or optimization and keeping the same parametrization.

Other biologically plausible alternatives that predict attention using neurodynamic modeling [184][82][67] do not provide a unified model of the visual cortex able to reproduce these distinct tasks simultaneously, and specifically, using real static or dynamic images as input. We suggest that V1 computations work as a common substrate for several tasks simultaneously. Although latest hypotheses about the SC confirm that saliency is processed in the SC and not by the visual cortex, corresponding to a distinct, feature-agnostic saliency map [336][352], we claim the importance of the mechanisms of V1 to be responsible for computing distinctiveness between the stated low-level features, which might conjunctively contribute to the generation of saliency [185][183][372]. However, modeling the computations of the pathways from the RGC to the SC would be of interest for a more integrated and complete model of eye-movement prediction, seeing the roles of the distinct projections to the SC and their computations, alternatively involved in the control of eye movements.

Visual Scanpaths and Relevance: Part III
Top-down attention



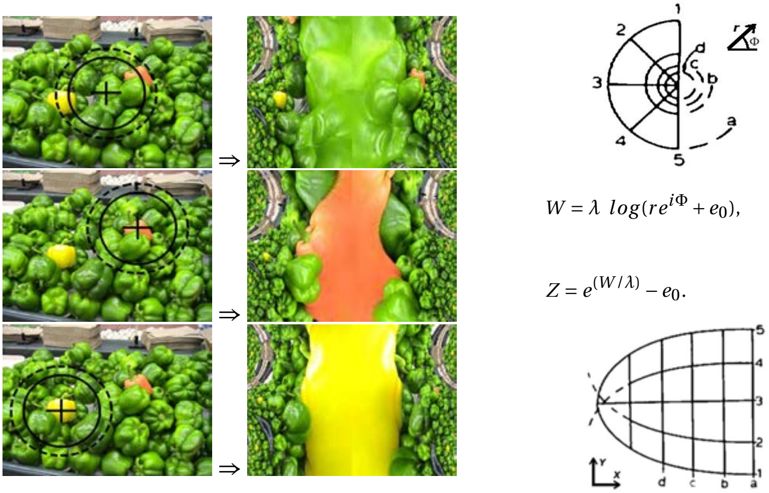
Extracting saccade sequences given any task.

5 NSWAM-CM: Cortical Magnification and Feedback Connections

Latest work from Berga & Otazu [26] has shown that the same model (without changing its parametrization) is able to predict saliency using real and synthetic color images. We propose to extend the model providing saliency computations with foveation, concerning distinct viewpoints during scene observation (mapping retinal projections towards V1 retinotopy) as a main hypothesis for predicting visual scanpaths. Furthermore, we also test how the model is able to provide predictions considering recurrent feedback mechanisms of already visited regions, as well as from visual feature and exemplar search tasks with top-down inhibition mechanisms.

5.1 Cortical Magnification mechanisms in the brain

The human eye is composed by RP but these are not homogeneously or equally distributed along the retina, contrarily to digital cameras. RP are distributed as a function of eccentricity with respect the fovea (or central vision)[303]. Fovea's diameter is known to comprise ~5deg of diameter in the visual field, extended by the parafovea (~5-9deg), the perifovea (~9-17deg) and the macula (~17deg). Central vision is known to provide maximal resolution at ~1deg of the fovea, whereas in periphery (~60-180- deg) there is lower resolution for the retinotopic positions that are further away from the fovea. These effects are known to affect color, shape, grouping and motion perception of visual objects (even at few degrees of eccentricity), making performance on attentional mechanisms as eccentricity-dependent [60]. Axons from the nasal retina project to the contralateral LGN, whereas the ones from the temporal retina are connected with the ipsilateral LGN. These projections [342] make the left visual field send inputs of the LGN towards the right hemifield of V1, similarly for the case of the right visual field to the left hemifield of V1. We have modeled these projections with a cortical magnification function [283][383, Section 2.3.1] using 128 mm of simulated cortical surface. The visual space is transformed to a cortically-magnified space (with its correspondence of millimeter for each degree of visual angle) with a logarithmic mapping function. The pixel-wise cartesian visual space is transformed to polar coordinates in terms of eccentricity and azimuth for a specific foveation instance, then transformed to coordinates in mm of cortical space (see an example in Figure 5.1).



$$W = \lambda \ log(re^{i\Phi} + e_0), \tag{5.1}$$

$$Z = e^{(W/\lambda)} - e_0. \tag{5.2}$$

Figure 5.1 – **Left:** Examples of applying the cortical magnification function (transforming the visual space to the cortical space) at distinct views. **Right:** Illustration of how polar coordinates (Z-plane) of azimuth $\Phi = (1, 2, 3, 4, 5)$ in the left visual field at distinct eccentricities $r = (d, c, b, a)$ are transformed to the cortical space (W-plane) in mm (X and Yi axis values), adapted from [283]. Equations 5.1 & 5.2 express the monopole direct and inverse cortical mapping transformations (parameters set as $\lambda = 12$mm and $e_0 = 1$deg [383, Section 2.3.1]).

Acknowledging that the visual space for digital images is represented with either a squared or rectangular shape, we computed the continuation of cortical coordinates by symmetrically mirroring existing coordinates of the image with their correspondence of visual space outside boundaries in the cortical space. In that manner, we exclude possible effects of zero-padding over recurrent processing while preserving 2D shapes for our feature representations. For this case, these responses can be minimized by the inverse and repeating the same process at specific interaction cycles. Schwartz's mapping has been applied over the wavelet coefficients represented in Figure 4.2, as basis functions are convolved in the visual space, later magnified to the cortical space for representing V1 signals. These signals will serve as input to excitatory pyramidal cells, projected to their respective iso-orientation domains at distinct RF sizes.

Table 5.1 – Overview of the model, following Nordlie et. al.'s format [215]. Further explanation for model variables and parameters is in [239, Supporting Information S1].

A	**Model Summary**
Populations	Four: excitatory (x), inhibitory (y)
Topology	–
Connectivity	Feedforward: one-to-all, Feedback: one-to-all, Lateral: all-to-all (including self-connections)
Neuron model	Dynamic rate model
Channel models	–
Synapse model	Piece-wise linear synapse
Plasticity	–
Input	External current in lower (I) or higher (I_c) cortical areas and random noise (I_0)
Measurements	Firing-rate (g_x and g_y)

B	**Populations**	
Name	**Elements**	**Size**
x	Sigmoidal-like neuron	$K_x = M \times N \times \Theta \times S = 64 \times 128 \times 3 \times 8$
y	Sigmoidal-like neuron	$K_y = K_x$

C	**Connectivity**		
Name	**Source**	**Target**	**Pattern**
J_{xx}	x	x	Excitatory, toric, all to all, non-plastic
J_0	x	x	Excitatory, constant $J_0 = 0.8$
W_{xy}	x	y	Inhibitory, toric, all to all, non-plastic
W_{yx}	y	x	Inhibitory, toric, all to all, non-plastic

D	**Neuron and Synapse Model**
Name	V1 neuron
Type	Dynamic rate model
Synaptic dynamics	$J_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.126e^{(-\beta/d_s)^2-2(\beta/d_s)^7-d_s^2/90} \quad (5.3)$ $W_{[is\theta,js'\theta']} = \lambda(\Delta_s)0.14(1-e^{-0.4(\beta/d_s)^{1.5}})e^{-(\Delta_\theta/(\pi/4))^{1.5}} \quad (5.4)$
Membrane potential	$\dot{x}_{is\theta} = -\alpha_x x_{is\theta} - g_y(y_{is\theta}) - \sum_{\Delta_s,\Delta_\theta\neq 0} \psi(\Delta_s,\Delta_\theta)g_y(y_{is}+\Delta_{s\theta}+\Delta_\theta) + J_0 g(x_{is\theta}) + \sum_{j\neq i,s',\theta'} J_{[is\theta,js'\theta']}g_x(x_{js'\theta'}) + I_{is\theta} + I_0, \quad (5.5)$ $\dot{y}_{is\theta} = -\alpha_y y_{is\theta} - g_x(x_{is\theta}) + \sum_{j\neq i,s',\theta'} W_{[is\theta,js'\theta']}g_x(x_{js'\theta'}) + I_c \quad (5.6)$

E	Input
Type	**Description**
Sensory (bottom-up)	Input to excitatory neurons, $I^t_{i;o} = \omega^t_{iso\theta}$
Control (top-down)	Input to inhibitory interneurons, $I_c = 1.0 + I_{noise} + I_{vs} + I_{ior}$

F	Measurements
Mean Firing-rate of excitatory neurons for τ=10 membrane time ($M(\omega^{p=[+,-]}_{iso\theta})$).	

After integrating the cortical mapping, the network of NSWAM-CM is in total composed of 1.18 × 6 neurons (accounting for 3 opponent channels, both ON/OFF polarities and RF sizes of 128 × 64 × 3×8).

5.2 Generating Saccade Sequences

We have defined the higher active neurons (Equation 5.7) as the locations for saccades in the visual space (i,j) by decoding the inverse of the cortical magnification (Equation 5.2) of their respective retinotopic position ("i" neuron at X,Yi).

$$MAX_W(X, Yi) = argmax(\hat{S}) \rightarrow MAX_Z(r, \Phi) \rightarrow MAX_V(i, j), \tag{5.7}$$

In Figure 5.2 we schematize the pipeline of the model, summarized in 8 procedures:

1. Transform RGB components to the CIE Lab space. [Chapter 4.1.1]
2. Extract orientation features at distinct scales with the DWT. [Chapter 4.1.2]
3. Feature maps are foveated with a log-polar cortical mapping transform (initial fixation set by default at the center of the image). [Chapter 5.1]
4. Firing rates are obtained with neurodynamical model of horizontal connections in V1. [Chapter 4.2]
5. Firing rates from feature maps are integrated to a unique map (with a inverse transform from previous methods) to select next fixation point. [Chapter 4.3][Chapter 5.2].
6. Activity from highest retinotopic position is set as an Inhibition of Return mechanism for future fixations (step 3). [Chapter 5.3.2]
7. Top-down selection can be introduced to the model as an inhibitory signal from coefficients of learned categories [Chapter 5.3.1]

8. Memory formations have been learned from wavelet coefficient statistics (chromatic opponency, scale and orientation), given images and segmentation masks [Chapter 5.3.1]

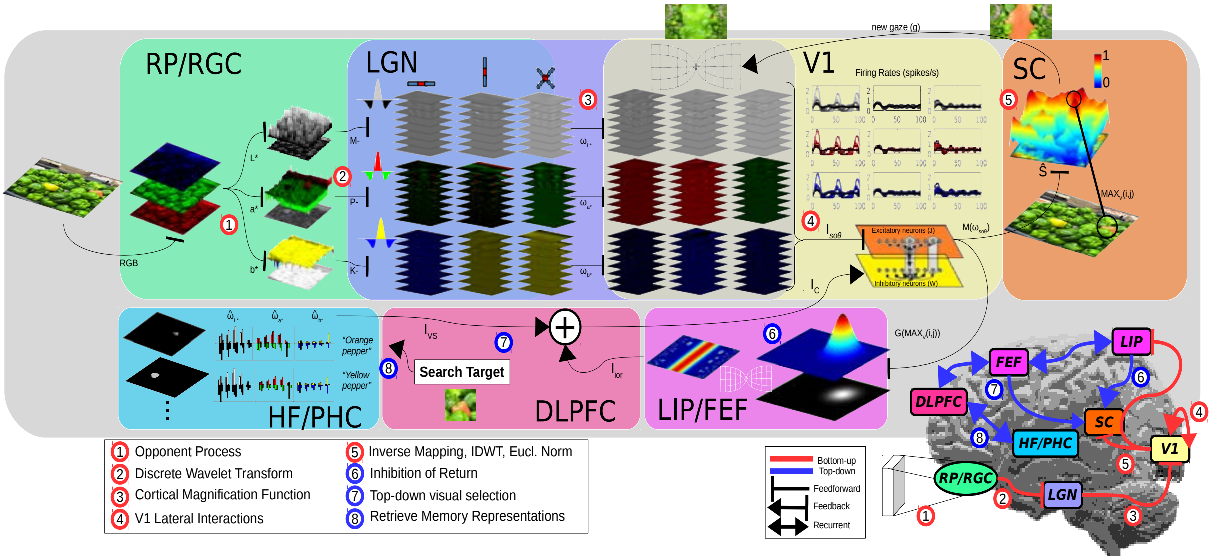


Figure 5.2 – Diagram illustrating how visual information is processed by NSWAM-CM, including a brain drawing of each bottom-up and top-down attention mechanisms and their localization in the cortex (bottom-right).

5.3 Attention as Top-down inhibition

An additional purpose of our work is model of attentional mechanisms beyond pre-attentive visual selection. Instead of analyzing the scene serially, the visual brain uses a set of attentional biases to recognize objects, their relationships and their importance with respect the task, all given in a set of visual representations. Similarly to the saliency map, the priority map can be interpreted as a unique 2D representation for eye movement guidance formed in the SC, here including top-down (not guided by the stimulus itself) and recurrent information as visual relevance. This phenomena suggests that executive, long-term and short-term/working memory correlates also direct eye movement control[173, 245, 350]. Previous hypotheses model these properties by forming the priority map through selective tuning [326, 327]. Selective tuning explains attention mechanisms as a hierarchy of winner-take-all processes. This hypothesis suggests that attention is focused with spatial inhibition at each layer of processing, creating selection, restriction and suppression through inhibition of certain locations, task-irrelevant cues and previously seen locations. Latest hypotheses [4] confirm that striate cortical activity gain can be modulated by SC responses, additional modulations arise from pulvinar to extrastriate visual areas. By simulating the formetly mentioned executive and recurrent activity as top-down inhibition in our model, we are able to perform task-specific visual selection (VS) and inhibition of return (IoR) mechanisms.

5.3.1 Top-down selection

Goal-directed or memory-guided saccades imply executive control mechanisms that account for task requirements during stimulus perception. The dorsolateral prefrontal cortex (DLPFC) is known to be responsible for short-term spatial memory, to retrieve long-term memory signals of object representations (through projections towards the para- and hippocampal formations) as well as to perform reflective saccade inhibition, among other functions. These inhibitory signals, later projected to the frontal eye field (FEF) are able to direct gaze during search and smooth pursuit tasks [243, 245, 278] (also suggested to be crucial for planning intentional or endogenously-guided saccades), where its signals are sent to the SC. Latest studies [372] put forward that V1 influences both saliency and top-down learning during visual detection tasks. By feeding our model with inhibitory signals (I_c shown in Figure 4.3 and Table 5.1E) we can simulate top-down selection at each level of visual processing. In this case, a new term $I_{\{vs\}}$ is added to the top-down inhibition of our cortical signals that will be projected to the SC during each gaze (similarly to a boolean selection of specific features [134]). This term is modulated with a constant factor $\alpha_{\{vs\}}$. In this implementation, we can perform distinct search tasks such as feature search (by manually selecting the features, or selecting features with maximal responses, as Ω=$MAX_{p,s,o,\theta}(\omega)$), exemplar and categorical object search (by processing the mean of responses $\hat{\omega}$ from wavelet coefficients of a single or several image samples. These would serve as RF activations to be stored as weights in our low-level memory representations) that will be used as inhibitory modulation for the task execution:

$$I_{\{vs\}} = \alpha_{\{vs\}} \cdot \begin{cases} MAX_{p,s,o,\theta}(\omega) & \text{,feature-selective} \\ (\sum_{i=1}^{N} \omega_{pso\theta})/N & \text{,category-specific} \end{cases} \tag{5.8}$$

5.3.2 Inhibition of Return

During scene viewing, saccadic eye movements show distinct patterns of fixations [89], directed by exploratory purposes or either towards putting the attentional focus on specific objects in the scene. For the former case, the HVS needs to ignore already visited regions (triggering anti-saccades away from these memorized regions) during a period of time before gazing again towards them. This phenomena is named inhibition of return [106], and similarly involves extracting sensory information and short-term memory during scene perception. As mentioned before, DLPFC is responsible of memory-guided saccades, and this function might be done in conjunction with the parietal cortex and the FEF. The parietal areas (LIP and PEF)[38, 243, 245] are known to be responsible of visuospatial integration and preparation of saccade sequences. These areas conjunctively interact with the FEF and DLPFC for planning reflexive visually-guided saccades. Acknowledging that LIP receives inputs from FEF and DLPFC, the role of each cannot be disentangled as a unique functional correlate for the IoR. Following the above and similarly to the previous section, we have modeled return mechanisms as top-down cortical inhibition accounting for previously-viewed saccade locations. Thus, we added an inhibition input $I_{\{IoR\}}$ at the start of each saccade, which will determine our IoR mechanism:

$$\begin{aligned} I_{\{IoR\}}^{g,t=0} &= argmax(\hat{S})\mathring{\text{u}}G(MAX_V(x,y)) + I_{\{IoR\}}^{g-1}, \\ I_{\{IoR\}}^{g,t>0} &= \alpha_{\{IoR\}}(I_{\{IoR\}}^{t-1}) \prod_{i=1}^{10\tau} e^{log(\beta_{\{IoR\}})/\tau}. \end{aligned} \tag{5.9}$$

This term is modulated with a constant power factor $\alpha_{\{IoR\}}$ and a decay factor $\beta_{\{IoR\}}$, which in every cycle will progressively reduce inhibition. The spatial region of the IoR has been defined as a gaussian function centered to the previous gaze (g), with a spatial standard deviation dependent on a specific spatial scale and a peak with an amplitude of the maximal RF firing rate. Inhibitory activity is accumulated to the same map and can be shown how is progressively reduced during viewing time (Fig. 5.10). Alternatively illustrated in Itti et al.'s work [142], the IoR can be applied to static saliency models by substracting the accumulated inhibitory map to the saliency map during each gaze ($\hat{S} - I_{\{IoR\}}^{g}$).

5.4 Evaluation metrics for Scanpaths and Visual Search

Saliency metrics, largely explained by Bylinskii et al. [55], usually compare model predictions with human fixations during the whole viewing time, regardless of fixation order. In our study is also represented the evolution of prediction scores for each gaze. For the case of scanpaths, we evaluated saccade sequences by analyzing saccade amplitude (SA) and saccade landing (SL) statistics. These are calculated using euclidean distance between fixation coordinates (distance between saccade length for SA and distance between locations of saccades for SL).

Initial investigations on visual attention [320, 365] during visual search tasks formulated that reaction times of finding a target (defined in a region of interest/ROI) among a set of distractors are dependent on set size as well as target-distractor feature contrast. In order to evaluate performance on visual search, we utilised two metrics that account for the ground truth mask of specific regions for search and the saliency map (in this context, it could be considered as a "relevance" map) or predicted saccade coordinates (from locations with highest neuronal activity). We have used the SI metric for calculating how well a saliency (or object search) map falls inside a particular ROI. For the case of saccades in visual search, we considered to calculate the probability of fixations inside the ROI (PFI).

5.5 Results for predicting Saliency

Saliency predictions have been computed from initial biologically-inspired saliency models for comparison. Our model has been computed without (NSWAM) and with foveation (NSWAM-CM), as a mean of cortically-mapped saliency computations through a loop of 1, 2, 5 and 10 saccades. The tasks performed for these mostly consist of freely looking at each image during 5000 ms, looking at the "most salient objects" or searching for specific objects of interest. We have selected these datasets to evaluate prediction performance at distinct scene contexts. Indicators of psychophysical consistency of the models has been presented, evaluating prediction performance upon fixation number and feature contrast. Visual search performance has been evaluated by computing predictions of locating specific objects of interest.

Based on the shuffled metric scores, traditional saliency models such as AIM overall score higher on real scene images (Fig 5.3), scoring $sAUC_{AIM}$=.663, and $InfoGain_{IKN}$=.024. For the case of nature images (Fig 5.4), our non-foveated and foveated versions of the model (NSWAM and NSWAM-CM) scored highest on both metrics ($InfoGain_{NSWAM}$=.168 and $sAUC_{NSWAM-CM10}$=.567). As mentioned before, fixation center biases are present when the task and/or stimulus do not induce regions that are enough salient to produce bottom-up saccades. Nature scene images lack of semantic (man-made) information that might contribute to top-down guided eye movements and inter-participant differences, in contrast to real image scenes [140]. Adding to that, in real images (Toronto and KTH), which contain dense image representations, there are no concrete places for saliency. This phenomena is seemingly presented in our models' saliency maps from 1st to 10th fixations (Figs. 5.3-5.4, rows 5-8), where salient regions are presented to be less evident across fixation order.

Model	sAUC	InfoGain
Human Fix.	.904	2.42
IKN [142]	.649	-.024*
AIM [49]	.663*	-.579
NSWAM	.631	-.552
NSWAM-CM1	.636	-.818
NSWAM-CM2	.644	-.738
NSWAM-CM5	.650	-.701
NSWAM-CM10	.655	-.692

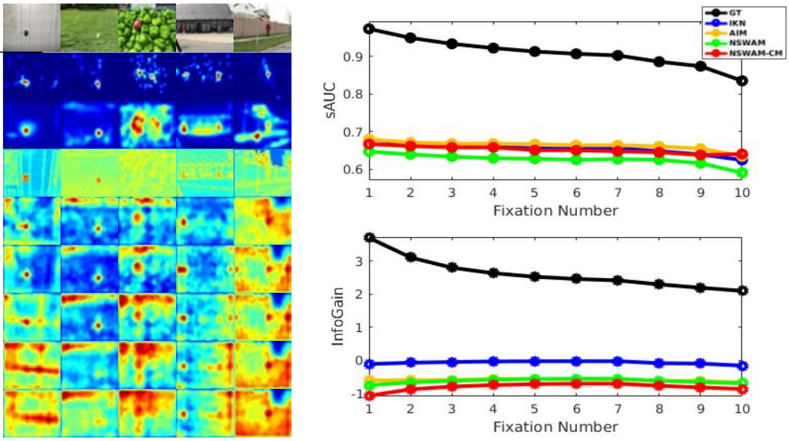


Figure 5.3 – Saliency metrics for Toronto (Bruce & Tsotsos [50]) Eye Tracking Dataset

Model	sAUC	InfoGain
Human Fix.	.822	1.41
IKN [142]	.551	-.172
AIM [49]	.552	-.509
NSWAM	.565	-.168*
NSWAM-CM1	.564	-.227
NSWAM-CM2	.566	-.213
NSWAM-CM5	.566	-.211
NSWAM-CM10	.567*	-.209

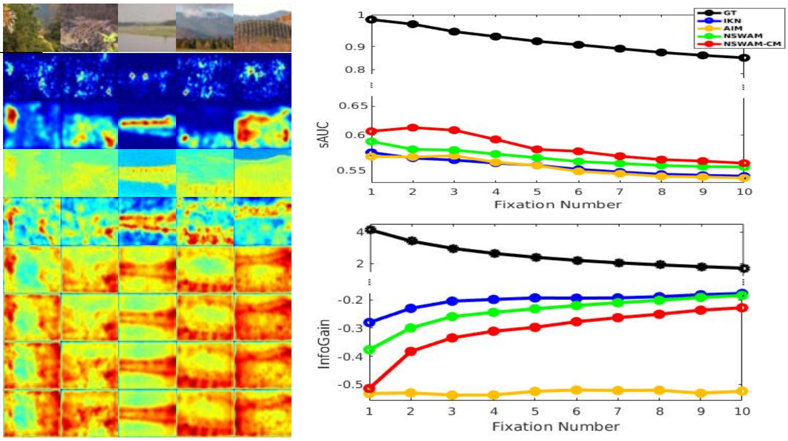


Figure 5.4 – Saliency metrics for KTH (Kootra et al'.s [160]) Eye Tracking Dataset

Model	sAUC	InfoGain
Human Fix.	.623	.777
IKN [142]	.562	-.724*
AIM [49]	.544	-6.49
NSWAM	.567*	-1.01
NSWAM-CM1	.561	-1.24
NSWAM-CM2	.563	-1.14
NSWAM-CM5	.565	-1.09
NSWAM-CM10	.567*	-1.07

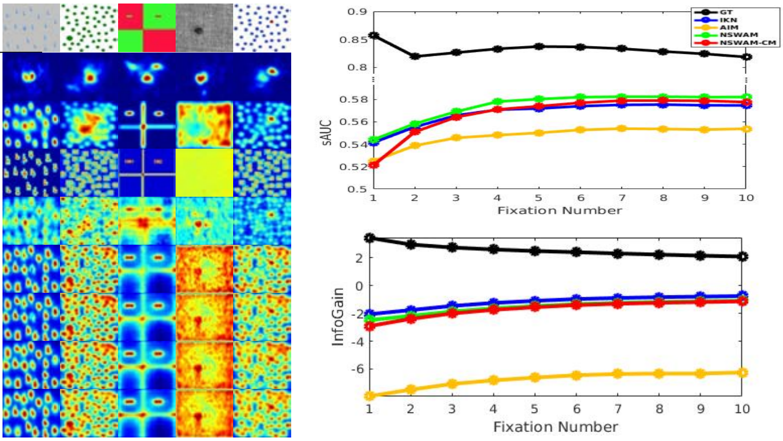


Figure 5.5 – Saliency metrics for $CAT2000_{Pattern}$ (Borji & Itti [42]) Dataset

Model	sAUC	InfoGain
Human Fix.	.860	2.80
IKN [142]	.608	-.233
AIM [49]	.557	-18.2
NSWAM	.622*	-.149
NSWAM-CM1	.617	-.204
NSWAM-CM2	.622*	-.164
NSWAM-CM5	.620	-.139
NSWAM-CM10	.618	-.131*

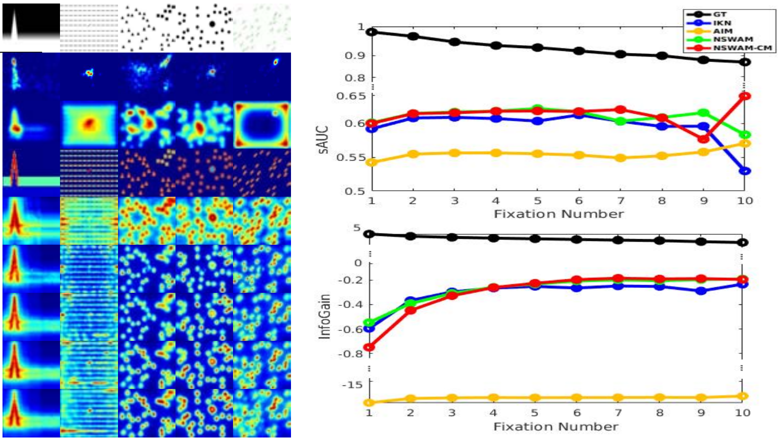


Figure 5.6 – Saliency metrics for SID4VAM (Berga et al. [22]) Eye Tracking Dataset

In synthetic image patterns ($CAT2000_P$), both of our model versions outperforms other models $sAUC_{NSWAM,NSWAM-CM}$=.567. Center biases are present in such dataset (see Fig. 5.5, "Human Fix." heatmaps), seemingly reproduced by IKN in the illustration ($InfoGain_{IKN}$=-.724). Quantitatively, these tendencies should not be likely to be considerable as indicators of saliency. Even if shuffled metrics try to penalize for these effects, systematic tendencies cannot be discarded from model evaluations (these are particular for each dataset task and stimulus properties). For the case of SID4VAM dataset (Fig. 5.6), salient regions are labeled with specific feature type and contrast, and fixation patterns present lower center biases (due to mainly being based a singleton search type of task with a unique salient target). Our model presents highest scores on both metrics ($sAUC_{NSWAM,NSWAM-CM2}$=.622 and $InfoGain_{NSWAM-CM10}$=-.131).

In Figs. 5.3-5.6 are compared the average score per gaze of human fixations and saliency model predictions. It can be observed that prediction scores for all models decrease as a function of gaze number. Scores of probability density distributions of human fixations (in comparison to fixation locations) decrease around 10% the sAUC after 10 saccades. This decrease of performance is not reproduced by any of the presented models, instead, most of them show a flat or slightly increasing slopes for the case of sAUC scores and logarithmically increasing scores for InfoGain. NSWAM and NSWAM-CM present similar results upon fixation number.

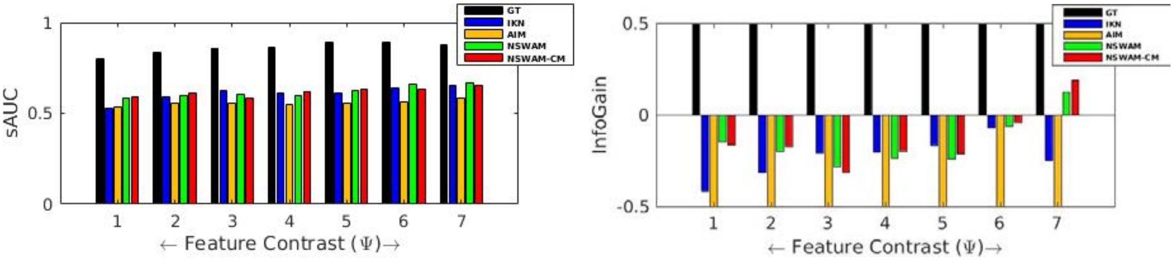


Figure 5.7 – sAUC and InfoGain scores for each relative target-distractor feature contrast

In SID4VAM, stimuli are categorized with specific difficulty (according to the relative target-distractor feature contrast). With these, we computed the score for each relative contrast instance (Ψ) in Fig. 5.7. Ideal conditions (following the Weber law) determine that if there is less difficulty for finding the salient region (higher target-distractor contrast), saliency will be focused on that region. Conversely, fixations would be distributed on the whole scene if otherwise. After computing each low-level stimulus instance with the presented models and evaluating results with the same metrics, our saliency model (NSWAM and NSWAM-CM) presents better performance than AIM and IKN while increasing score at higher feature contrasts.

Discussion

Quantitatively, systematic tendencies in free-viewing (center biases, inter-participant differences, etc.[306]) should not be likely to be considered as indicators of saliency. Although shuffled metrics try to penalize for these effects, benchmarks do not compensate for these tendencies from model evaluations (these are particular for each dataset task and stimulus properties). Acknowledging that first saccades determine bottom-up eye movement guidance [8, 381], it is a phenomenon also present in our experimental data (in terms of the decrease of performance with respect fixation region probability compared to fixation locations). In that aspect, evaluating first fixations with more importance could define new benchmarks for saliency modeling, similarly with stimuli where feature contrast in salient objects is quantified. Ideal conditions (following the Weber law) determine that if there is less difficulty for finding the salient region (higher target-distractor contrast), saliency will be focused on that region. Conversely, fixations would be distributed on the whole scene if otherwise. Our model presents better performance than other biologically-inspired ones accounting for these basis.

5.6 Results for predicting Scanpaths

Illustration of scanpaths from datasets presented in previous section were computed with scanpath models in Fig. 5.8. Scanpaths are predicted by NSWAM-CM during the first 10 saccades, by selecting maximum activity of our model for every saccade. We have plotted our model's performance in addition to Boccignone&Ferraro's, LeMeur&Liu's and STAR-FC predictions (Fig. 5.9). Saccade statistics show an initial increment of saccade amplitude, decreasing as a function of fixation number. Errors of SA and SL (ΔSA and ΔSL) are calculated as absolute differences between model predictions and human fixations. Values of ΔSL appear to be lower and similar for all models during initial fixations.

In our study is also represented the evolution of prediction scores for each gaze. For the case of scanpaths, we evaluated saccade sequences by analyzing saccade amplitude (SA) and saccade landing (SL) statistics. These are calculated using euclidean distance between fixation coordinates (distance between saccade length for SA and distance between locations of saccades for SL). Prediction errors are shown to be sustained or increasing for CLE and NSWAM-CM (maybe due to their lack of processing higher level features, experimental center biases, etc.). Errors on ΔSA predictions are lower for LeMeur&Liu's and STAR-FC models, retaining similar saccades (except for synthetic images of SID4VAM). Although these errors are representative in terms of saccade sequence, we also computed correlations of models' SA with GT (ρSA). In this last case, NSWAM-CM presents most higher

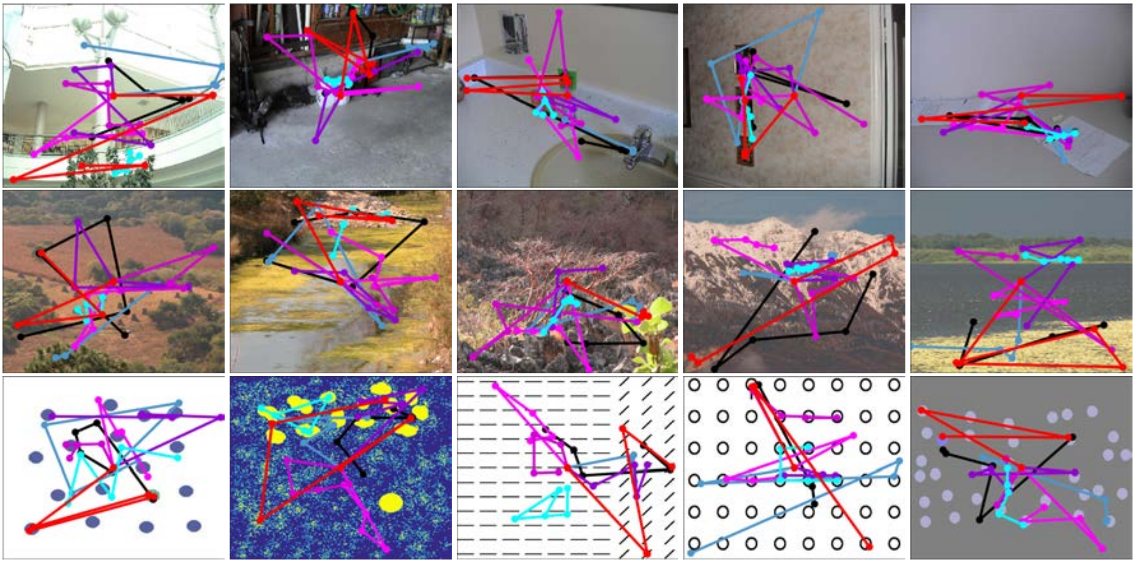


Figure 5.8 – Examples of visual scanpaths for a set of real (1st row), nature (2nd row) and synthetic (3rd row) images. Model scanpaths correspond to CLE [39], LeMeur$_{Natural}$, LeMeur$_{Faces}$, LeMeur$_{Landscapes}$ [202], STAR-FC[357] and NSWAM-CM (ours).

correlation values for Toronto, Kootstra and CAT2000$_P$ (ρSA$_{Toronto}$=-.38, p=.09; ρSA$_{KTH}$=.012, p=.96; ρSA$_{CAT2000_P}$=.28, p=.16) than other models. Most of them seem to accurately predict SA for SID4VAM (which contains mostly visual search psychophysical image patterns), with ρSA between .7 and .8.

Our scanpath model tend to predict eye movements with large mean saccade amplitudes $\{M(SA)_{Toronto}$=7.8±3.5; $M(SA)_{KTH}$=13±6.1; $M(SA)_{CAT2000_P}$=15.7±6.7; $M(SA)_{SID4VAM}$=15.7±6.9 deg$\}$, whereas human fixations combine both short and large saccades $\{M(SA)_{Toronto}$=4.6±1; $M(SA)_{KTH}$=6.7±.5; $M(SA)_{CAT2000_P}$=5.1±.9; $M(SA)_{SID4VAM}$=5.8±1.5 deg$\}$. In that aspect, our prediction errors might arise from not correctly predicting focal fixations.

We simulated the inhibition factor for all datasets by substracting the inhibition factor $I_{\{IoR\}}$ to our models' saliency maps (NSWAM+IoR). After computing prediction errors in SA and SL for a single sample (Fig. 5.11-Top), best predictions seem to appear at decay values of $\beta_{\{IoR\}}$ between .93 and .98, which corresponds to 1 to 5 saccades (similarly explained by Samuel & Kat [276] and Berga et al. [22],

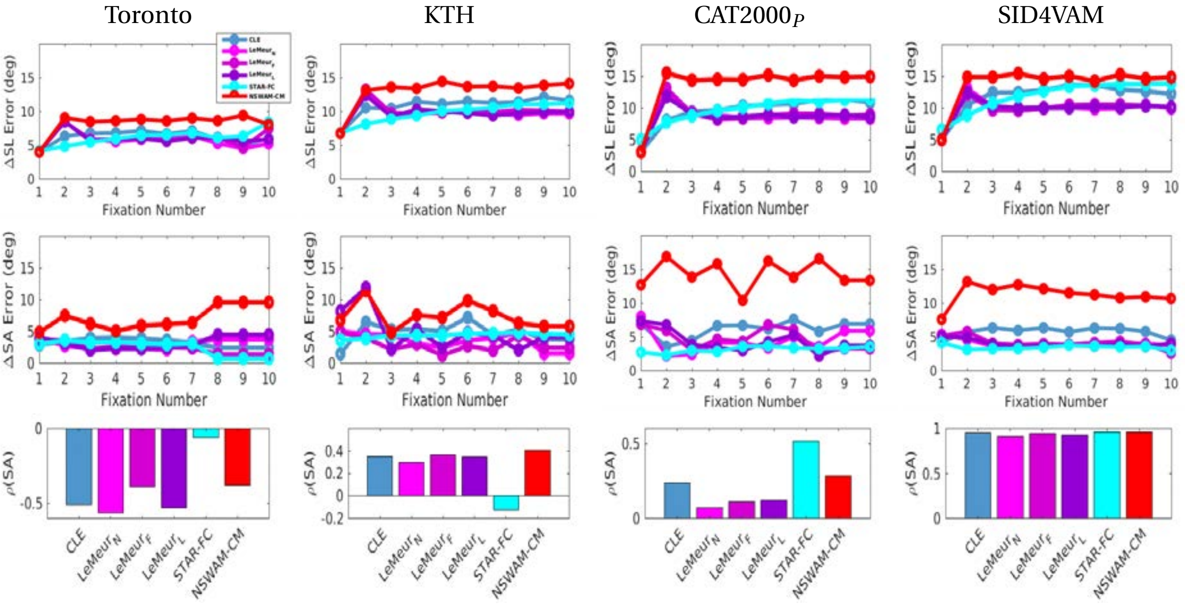


Figure 5.9 – **1st row:** Prediction errors in Saccade Landing (ΔSL) for real indoor/outdoor (Toronto), nature (KTH) and synthetic (CAT2000$_P$ and SID4VAM) image datasets. **2nd row:** Prediction errors in Saccade Amplitude (ΔSA) on same datasets. **3rd row:** Correlations of Saccade Amplitude (ρSA) with respect human fixations.

where takes from 300-1600 ms for the duration of the IoR, corresponding to 1 to 5 times the fixation duration). For the case of the $\sigma_{\{IoR\}}$, lowest prediction error (again, both in SA and SL) is found from 1 to 3 deg (in comparison, LeMeur & Liu [202] parametrized it by default as 2 deg). Results on ΔSA statistics have similar / slightly increasing performance until ($\beta_{\{IoR\}}$ <1) a single fixation time, decreasing at highest decay $\beta_{\{IoR\}}$ ≥5th saccade. For ΔSL values, errors in datasets such as KTH and SID4VAM are decreased at higher decay. For the latter, ΔSA errors are shown to decrease progressively at highest decay values ($\beta_{\{IoR\}}$ ≥.93). Lastly, when parametrizing the spatial properties of the IoR, saccade prediction performance is highest at lower size (with a near-constant error in SA and SL increasing about 1 deg for $\sigma_{\{IoR\}}$=1 to 8 deg on all datasets).

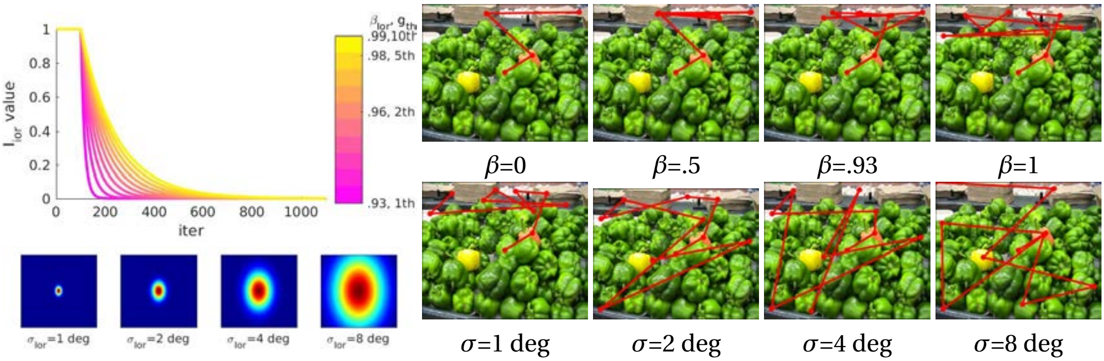


Figure 5.10 – **Left**: Evolution of inhibition factor for 100 mem.time (about 1000 iterations), corresponding approximately to performing 10 saccades to the model (top). Spatial representation of the IoR with distinct size (bottom). **Right**: Examples of scanpaths varying IoR decay factor (top, $\sigma_{\{IoR\}}$=2 deg, $\beta_{\{IoR\}}$={0,.5,.9,1}) or varying distinct IoR size (bottom, $\sigma_{\{IoR\}}$={1,2,4,8} deg,$\beta_{\{IoR\}}$=1).

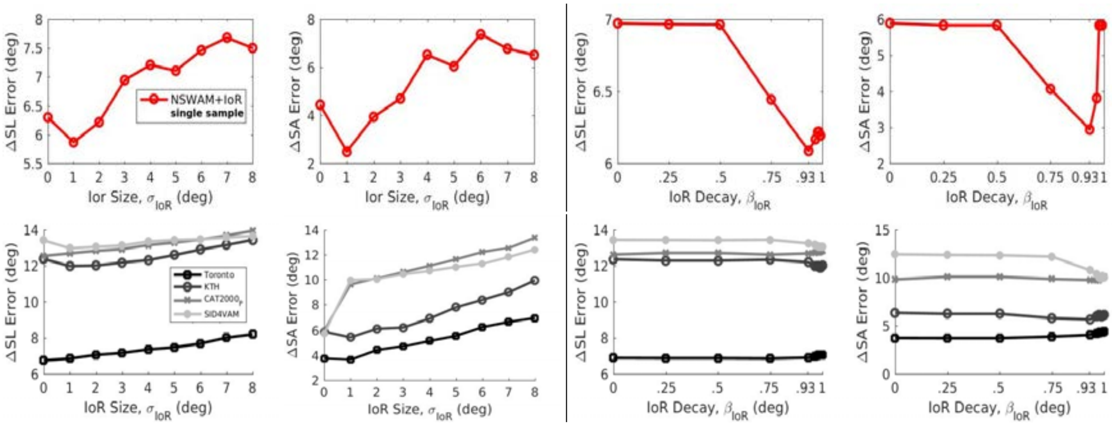


Figure 5.11 – Statistics of scanpath prediction (ΔSA and ΔSL) by the parametrization of IoR decay ($\beta_{\{IoR\}}$) and IoR size ($\sigma_{\{IoR\}}$) in a single sample (**top row**, from image scanpaths in Fig. 13) and saliency datasets (**bottom row**).

Discussion

Our model predictions on SA correlate better (i.e. obtain higher ρSA values) than other scanpath models (in terms of how SA evolves over fixations), however, prediction errors are higher in both SL and SA. We believe that these errors are caused by incorrectly predicting locations of fixations, but not for failing on predictions of the saccade sequence per se. These locations are mainly influenced by systematic tendencies in free-viewing (derived by center biases and/or focal fixations in a particular region of the image). Cortical magnification mechanisms might be responsible for processing higher saliency at regions outside the fovea, generating tendencies of uniquely capturing large saccades. These can be solved by processing high-level feature computations near the fovea, which would increase the probability of fixations at lower SA. We have to hesitate that first fixations are long known for being determinants of bottom-up attention [8, 22]. Instead, higher inter-participant differences [306] and center biases [269] increase as functions of fixation number, suggested as worse candidates for predicting attention. These parameters appear to specifically affect each stimuli differently (and accounting that each stimulus may convey specific semantic importance between each contextual element), which may relate to top-down attention but not to the image characteristics per se. We also want to stress the importance of foveation in our model. This is a major procedure for determining saccade characteristics (including oculomotor tendencies) and saliency computations, as it determines current human actions during scene visualization. The decrease of spatial resolution at increasing eccentricity provides the aforementioned properties, innate in human vision and invariant to scene semantics.

Adding an IoR mechanism has been seen to affect model activity and therefore scanpath predictions. In Fig. 5.10-Left we show how our inhibition factor ($I_{\{Ior\}}$) decreases over simulation time in relation to the parametrized decay $\beta_{\{IoR\}}$, as well as the projected RF size with respect the gaussian parameter $\sigma_{\{IoR\}}$. These variables (decay and size) affect either location of saccades and its sequence, modulating firing rate activity to already visited locations. It is shown in Fig. 5.10-Right that the initial saccade is focused on the salient region and then it spreads to a specific location in the scene, not repeating with higher value of inhibition decay or field size. In the next section we show how our model can preproduce eye movements beyond free-viewing tasks by modulating of inhibitory top-down signals.

5.7 Results on feature and categorical search

Comparison of results for NSWAM with bottom-up only and with top-down inhibition present higher scores for both SI and PFI (Fig. 5.12) using top-down inhibition (NSWAM+VS$_M$ and NSWAM+VS$_C$). Here, there is an increase of fixations inside the ROI: $\Delta(PFI)_{VS_M} \simeq$1% and $\Delta(PFI)_{VS_C} \simeq$6% for real object search and almost equal to saliency for synthetic image patterns, $\Delta(PFI)_{VS_M} \simeq$0% and

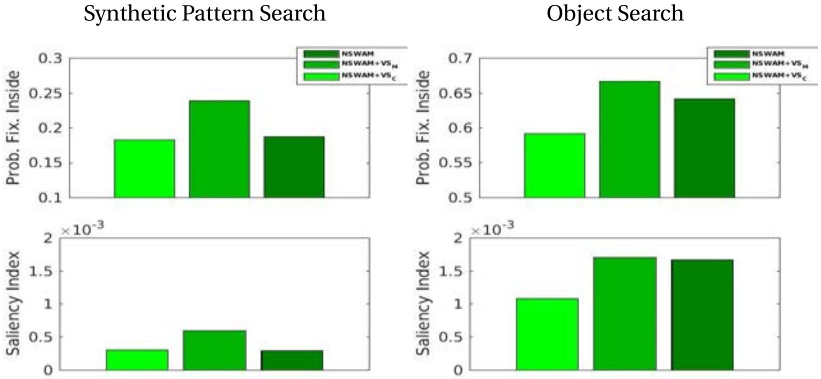


Figure 5.12 – Statistics of Saliency Index (**top row**) and Probability of Fixations Inside the ROI (**bottom row**) for synthetic image patterns (**left**) and salient object detection regions from real image scenes (**right**).

$\Delta(PFI)_{VS_C} \simeq$1%. The SI is also seen to increase for both cases, with differences of $\Delta(SI)_{VS_M}$=3.8×10^{-4} and $\Delta(SI)_{VS_C}$=5.9×10^{-4} for object search and $\Delta(SI)_{VS_M}$=3.1×10^{-4} and $\Delta(SI)_{VS_C}$=1.3×10^{-5} for psychophysical pattern search. Saliency metrics of sAUC and InfoGain (with Toronto's eye tracking dataset) increase with the search-based strategy {$\Delta(sAUC)_{VS_M}$=.018, $\Delta(sAUC)_{VS_C}$=.003; $\Delta(InfoGain)_{VS_M}$=.002, $\Delta(InfoGain)_{VS_C}$=.035}.

Free-viewing fixations are seemingly predicted with similar performance in comparison with NSWAM predictions (Fig. 5.3). Saliency metrics are similar or increasing with respect NSWAM for feature singleton search fixations {$\Delta(sAUC)_{VS_M}$=3.6×10^{-3}, $\Delta(sAUC)_{VS_C}$=2.9×10^{-3}; $\Delta(InfoGain)_{VS_M}$=4.1×10^{-2}, $\Delta(InfoGain)_{VS_C}$=9.4×10^{-4}}, but decrease for the case of free-viewing {$\Delta(sAUC)_{VS_M}$=-12×10^{-3}, $\Delta(sAUC)_{VS_C}$=-8.7×10^{-3}; $\Delta(InfoGain)_{VS_M}$=-13.7×10^{-2}, $\Delta(InfoGain)_{VS_C}$=-3.3×10^{-2}}.

We illustrated results of PFI and SI (Fig. 5.15) in relation to relative target-distractor feature contrast for cases of Brigthness, Color and Size differences, as well as the Set Size for searching a certain target patterns (i.e. a circle superposed by an oriented bar). After computing SI for each distinct psychophysical stimuli, we can see in Figs. 5.14-5.15 that our model performs best for searching differences with stimuli where there are differences in brightness, color, size and/or superimposed singletons, rather than for the case of different combination of orientations, specially with heterogeneous, nonlinear or categorical angle configurations.

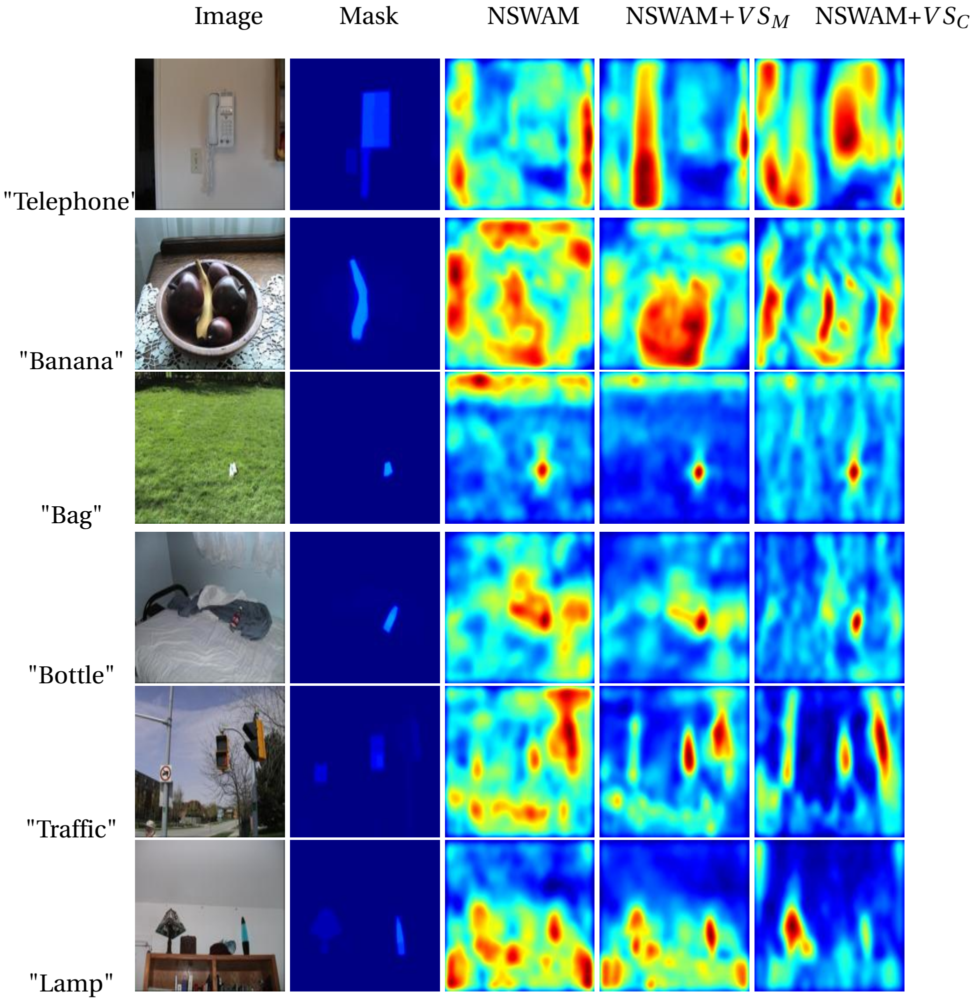


Figure 5.13 – Search instances with a specific ROI (Mask) based on a category/word exemplar.

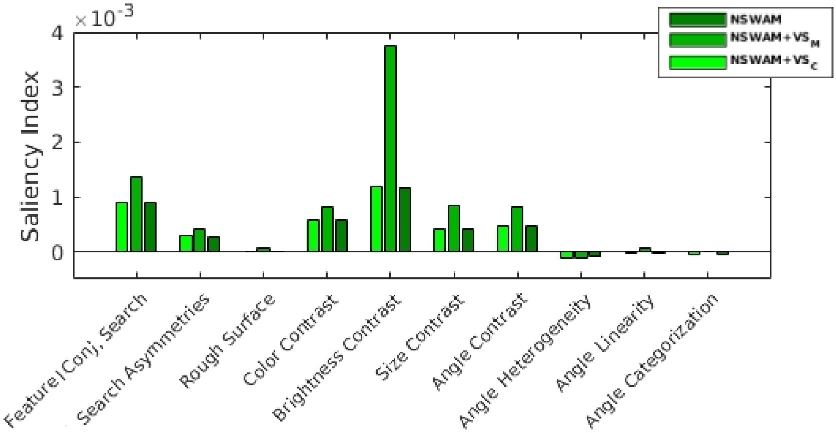


Figure 5.14 – Performance on visual search evaluated on each distinct low-level feature, stimulus instances are from SID4VAM's dataset [23].

Discussion

Overall results show that features computed by the top-down approach seemingly performs better in visual search than saliency, both considering features with maximal cortical activity (NSWAM+VS_M) and average statistics of low-level features (NSWAM+VS_C). When searching real objects, results in SI are higher for NSWAM+VS$_C$ (considering that dataset ROIs are selected from objects that are already salient). We suggest that considering scene statistics perform better when searching contextually complex exemplars. Here the combination of features could be implicit when processing image ROI average characteristics but not when using maximal activations, qualitatively shown in Fig. 5.13. The fact that SI scores are lower for free-viewing tasks in pop-out stimuli might be caused from influences of the center bias, presenting more fixations near the center in free-viewing [22]. Search in psychophysical image patterns is significatively more efficient in SI when selecting maximal feature activations (NSWAM+VS_M). Regarding that aspect, exemplar and categorical search for objects in real image scenes would require computations with a higher number of features [32, 198] (which would represent in more detail each cortical cell sensitivity).

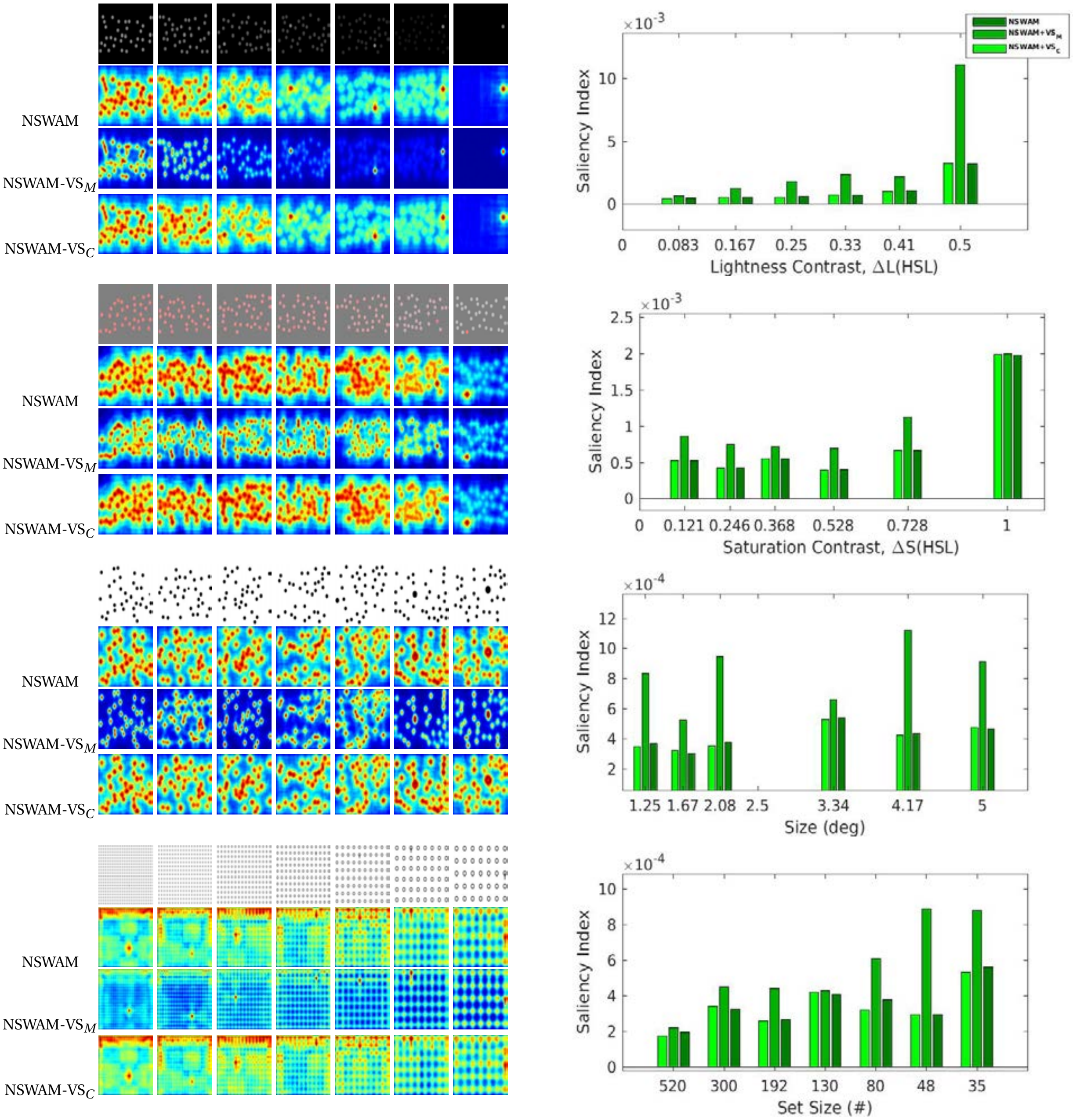


Figure 5.15 – Performance on visual search examples with a specific low-level feature contrast (for Brightness, Color or Size) and Set Size. We represented 7 instances ordered by search difficulty of each feature sample.

General Discussion

Current implementation of our V1 model is based on Li's excitatory-inhibitory firing rate network [184], following previous hypotheses of pyramidal and interneuron connectivity for orientation selectivity in V1 [105, 347]. To support and extend this hypothesis, distinct connectivity schemas (following up V1 cell subtype characterization) [114, 178] could be tested (e.g. adding dysynaptic connections between inhibitory interneurons) to better understand V1 intra-cortical computations. Furthermore, modeling intra-layer interactions of V1 cells [294] could explain how visual information is parallely processed and integrated by simple and complex cells[198], how distinct chromatic opponencies (P-,K- and M-) are computed at each layer [148], and how V1 responses affect SC activity (i.e. from layer 5) [214]. Testing contributions of each of these chromatic pathways (at distinct single/double opponencies and polarities), as well as distinct fusion mechanisms regarding feature integration, would define a more detailed description of how visual features affect saliency map predictions.

Previous and current scanpath model predictions could be considered to be insufficient due to the scene complexity and numerous factors (such as the task specificity, scene semantics, etc.) simultaneously involved in saccade programming. These factors increase overall errors on scanpath predictions, as systematic tendencies increase over time[22, 88, 269, 306], making late saccades difficult to predict. In that aspect, in free-viewing tasks (when there is no task definition), top-down attention is likely to be dependent on the internal state of the subject. Further understanding of high level attentional processes have only been approximated through statistical and optimization techniques with fixation data. It has also been later observed that fixations during free-viewing and visual search have distinct temporal properties. This could explain that saliency and relevance are elicited differently during viewing time. Latest literature on that aspect, discern two distinct patterns of fixations (either ambient or focal) where subjects first observe the scene (possibly towards salient regions), then focus their attention on regions that are relevant to them[89], and these influences are mainly temporal. Its modelization for eye movements in combination with memory processing is still under discussion. Current return mechanisms have long been computed by inhibiting the regions of previous fixations (spatially-based), nonetheless, IoR could also have feature-selective properties[133] to consider.

We suggest that not all fixations should have the same importance when evaluating saliency predictions. Nature and synthetic scene images lack of semantic (man-made) information, which might contribute to the aforementioned voluntary (top-down guided) eye movements [140]. Acknowledging that objects are usually composed by the combination of several features (either in shape, color, etc.), we should analyze if low-level features are sufficient to perform complex categorical search tasks. Extrastriate computations could allow the usage of object representations at higher-level processing, introducing semantically-relevant information and several image samples per category. Cortical processing of extrastriate areas (from V2 and V3) towards temporal (V4 & IT) and dorsal (V5 & MT) pathways [348, Section II][294] could represent cortical activity at these distinct levels of processing, modeling in more detail the computations within the two-stream hypothesis (what & where pathways). Color, shape and motion processing in each of these areas could generate more accurate representations of SC activity[350], producing more complex predictions such as microsaccadic and smooth pursuit eye movements.

Future Work

Current and future implementations of the model are able to process dynamic stimuli as to represent attention using videos. By simulating motion energy from V1 cells and MT direction selective cells [383, Section 2.3.5], would allow our model to reproduce object motion and flicker mechanisms found in the HVS. Moreover, foveation through more plausible cortical mapping algorithms [282] could provide better spatial detail of the cortical field organization of foveal and peripheral retinotopic regions and lateralization, currently seen to reproduce V1/V2/V3 physiological responses. Adding to that, hypercolumnar feature computations of geniculocortical pathways could be extended with a higher number of orientation and scale sensitivities with self-invertible 2D Log-Gabor filters [95]. In that regard, angle configuration pop-out effects and contour detection computations [10, 11] can be done by changing neuron connectivity and orientation tuning modulations.

We aim in future implementations to model the impact of feedback in cortico-cortical interactions with respect striate and extrastriate areas in the HVS. Some of these regions project directly to SC, including the intermediate areas (pulvinar and medial dorsal) and basal ganglia[243, 245, 350]. Our current implementation can be extended with a large scale network of spiking neurons [146], also being able to learn certain image patterns through spike-timing dependent plasticity mechanisms[85]. With such a network, the same model would be able to perform both psychophysical and electrophysiological evaluations while providing novel biologically-plausible computations with large scale image datasets.

Conclusion

In this study we have presented a biologically-plausible model of visual attention by mimicking visual mechanisms from retina to V1 using real images. From such, computations at early visual areas of the HVS (i.e. RP, RGC, LGN and V1) are performed by following physiological and psychophysical characteristics. Here we state that lateral interactions of V1 cells are able to obtain real scene saliency maps and to predict locations of visual fixations. We have also proposed novel scanpath computations of scene visualization using a cortical magnification function. Our model outperforms other biologically inspired saliency models in saliency predictions (specifically with nature and synthetic images) and has a trend to acquire similar scanpath prediction performance with respect other artificial models, outperforming them in saccade amplitude correlations. The aim of this study, besides from acquiring state-of-the-art results, is to explain how lateral connections can predict visual fixations and how these can explain the role of V1 in this and other visual effects. In addition, we formulated projections of recurrent and selective attention using the same model (simulating frontoparietal top-down inhibition mechanisms). Our implementation of these, included top-down projections from DLPFC, FEF and LIP (regarding visual selection and inhibition of return mechanisms). We have shown how scanpath predictions improve by parametrizing the inhibition of return, with highest performance at a size of 2 deg and a decay time between 1 and 5 fixations. By processing low-level feature representations of real images (considering statistics of wavelet coefficients for each object or feature exemplar) and using them as top-down cues, we have been able to perform feature and object search using the same computational architecture. Two search strategies are presented, and we show that both the probability to gaze inside a ROI and the amount of fixations inside that ROI increase with respect saliency. In previous studies, the same model has been able to reproduce brightness [239] and chromatic [65] induction, as well as explaining V1 cortical hyperexcitability as a indicator of visual discomfort [241]. With the same parameters and without any type of training or optimization, NSWAM is also able predict bottom-up and top-down attention for free-viewing and visual search tasks. Model characteristics has been constrained (in both architecture and parametrization) with human physiology and visual psychophysics, and can be considered as a simplified and unified simulation of how low-level visual processes occur in the HVS.

6 Modeling task on attention (Ongoing Work)

6.1 Visual Priming

Presenting a stimulus can facilitate or inhibit processing of subsequent stimulus, this phenomena is called perceptual priming. Likewise, task priming refers to the change of perceptual and cognitive outcome when observing a stimulus after presenting a task. This is easily observable in the "Monkey Business Illusion", [1] also named "change blindness" or "inattentional blindness" [293], where presenting the task "Count how many times the players wearing white pass the basketball" makes most participants unable to detect a gorilla walking in front of the scene. Alfred Yarbus' seminal work [373][Figure 6.1] explained these effects for the case of eye movement control. Both task and perceptually-relevant information interplay as effects biasing image recognition and decision-making [103, 155, 375][56, Chapter 3.12]. On that regard, our HVS modulates visual attention, learning and visual working memory [59, 162] upon previous knowledge [329, 354]. Acknowledging that stimulus have both visual and semantical characteristics, conscious and unconscious percepts might be facilitated or inhibited (becoming supraliminal or subliminal) by certain image characteristics [14, 253]. Our attentional mechanisms select either the regions or features of the scene based on the task to perform [117, Chapter 7], the context (visual and semantic relationships) [12, 35, 127, 224, 229][166, Chapter 8] and prior knowledge [329, 354]. Acknowledging that while viewing a scene there is both saliency (bottom-up) and relevance (top-down), selection for locations of fixations need to be biased upon the aforementioned priors. The concept of priming can be included as a mechanism of Selective Tuning hypothesis [326][325, Chapter 7], such modeling will be presented in this chapter.

[1] Inattentional blindness: https://www.youtube.com/watch?v=vJG698U2Mvo

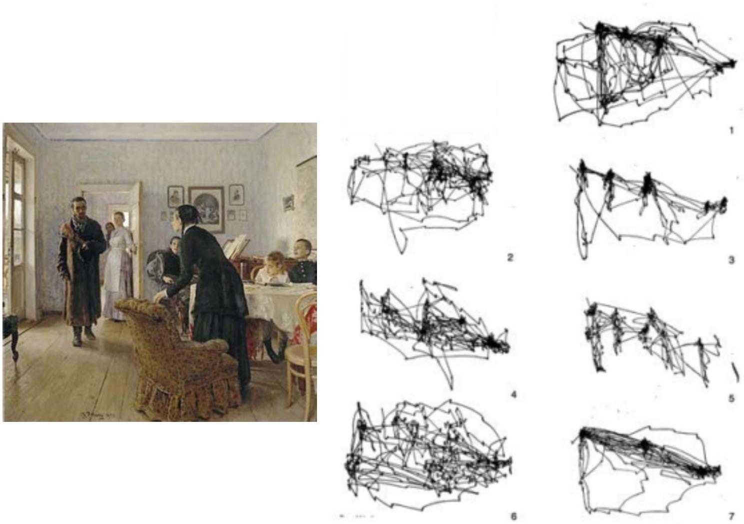


Figure 6.1 – *Unexpected Visitors* painting, by Ilya Repin. This picture elicits eye movement patterns according to distinct task specifications, represented as 1."Free examination"; 2."Estimate the material circumstances of the family in the picture"; 3."Give the ages of the people"; 4."Surmise what the family had been doing before the arrival of the 'unexpected visitor'"; 5."Remember the clothes worn by the people"; 6."Remember the position of the people and objects in the room"; 7."Estimate how long the 'unexpected visitor' had been away from the family". From [373].

6.2 Objectives

In this work we propose to model an architecture to enable tuning of attentional processing mechanisms to adapt vision given specific task instructions. We plan to provide a computational basis of processing object representations and locations for fixations according to stimulus characteristics and task-specific priming [52, 308, 373]. Navalpakkam & Itti [213] defined one of the first architectures that combine both visual and symbolic representations for attention. Later models [153] used "Bag of Words" of Shape and Color representations for learning class attention maps. Although its sound approach on predicting several tasks simultaneously, it was not yet implemented to work with real images, and its mechanisms are not inspired on a biological basis. For that, the task definition is to be integrated with the Selective Tuning Attentive Reference (STAR) Model [Figure 6.2-Left][327], presented to work for predicting saccade sequences with real image datasets [357]. The objectives for this model include:

1. To provide low- and high-level feature basis for the visual memory representations
2. To define a hierarchical lexico-semantic memory explaining the task specification
3. Determine influences of task over fixations (including memory, covert and overt attention)

6.3 STAR-FCT: Selective Tuning Attentive Reference - Task-based Fixation Controller

Our proposed implementation of methods in Long Term Memory (mLTM), Task Working Memory (tWM) and Visual Task Executive (vTE) [Figure 6.2-Right] are inspired by the Cognitive Programs architecture [328][325] by extending STAR Fixation Controller [327, 357], proposing functionally plausible definitions for reproducing human attentional processes in the brain.

6.3.1 Symbolic Representations: Understanding the Task

The task can be processed as a symbolic graph regarding lexical and semantical characteristics of the sentence, determining which objects, actions and locations are relevant and their importance with respect to the task. Here we integrated an ontology-based semantic similarity procedure from Wu & Palmer [370] that relates words with "is a" relationships in a common taxonomy (similarly to Navalpakkam & Itti's model [213]), defined by:

$$\begin{aligned} WUP(a,b) &= \frac{2 * depth(LCS(a,b))}{depth(a)+depth(b)}, \\ depth(a) &= min_{path}(root,a), \end{aligned} \tag{6.1}$$

where "a" and "b" are the words to compare, the "depth" is defined as the distance of the word with respect to the taxonomy "root" and the "LCS" or Least Common Subsummer is the most concrete taxonomical ancestor that subsumes both terms. Semantic similarity measures [199, 285][Equation 6.1] can relate learned words of a particular model (i.e. classification weights for categories/classes) with respect the words specified on the task (here using WordNet database [204]). This procedure

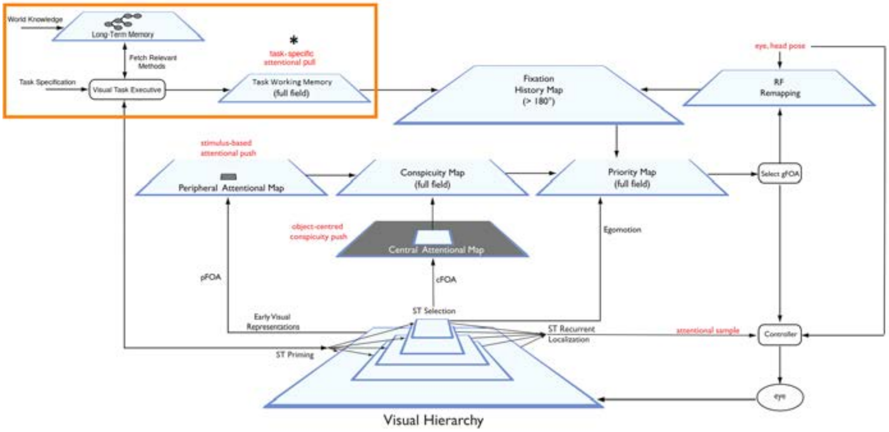

World Knowledge
Long-Term Memory
*
task-specific attentional pull
Fetch Relevant Methods
Task Specification
Visual Task Executive
Task Working Memory (full field)
eye, head pose
Fixation History Map (> 180°)
RF Remapping
stimulus-based attentional push
Peripheral Attentional Map
Conspicuity Map (full field)
Priority Map (full field)
Select gFOA
object-centred conspicuity push
Central Attentional Map
Egomotion
pFOA
cFOA
ST Selection
Early Visual Representations
ST Recurrent Localization
attentional sample
Controller
ST Priming
eye
Visual Hierarchy

Figure 6.2 – Illustration of STAR architecture including the integration of task definition modules (*). Here the image is blurred and cropped using a retinal transform [345], processed by a hierarchy of features representing low-level (Peripheral map) and high-level saliency (Central map). Fixations are predicted from WTA-like mechanisms from the Priority Map, which depends on previous fixations (Fixation History Map) and other Working Memory biases.

can able to make the model generalize to some extent to any class (whether or not exists in training data), abling to weight the map for the semantically-closest class with respect each one of the task.

6.3.2 Visual Representations: Low- and High-level features

Low- and high-level visual feature representations can be obtained both by computing features given the selected task categories. The STAR-FC scanpath model [357] uses low-level saliency from AIM [50] for computing peripheral map and high-level saliency from SALICON [135]) for computing the central map. Our aim is to define a set of methods that would allow to tune the hierarchical representations of peripheral and central attentional maps respectively. On the one hand, low-level features can be processed from sparse dictionaries (e.g. feature basis from LogGabors [95], ICA from features that maximize visual information [50], features that resemble V1 receptive fields [260], AlexNet first convolutional layers [96], etc.), being tuned using sparse coding [94, 191]. On the other hand, high-level features can be obtained from class activations from distinct layers of convolutional neural network models [387] (in that regard, previous STNet [36] and Priming Neural Networks [265] have already shown to work in the context of object localization, detection and segmentation). Our current implementation of high-level features has been performed using deep CNNs (e.g. VGG16, VGG19, ResNet50, etc.) pretrained on ImageNet.

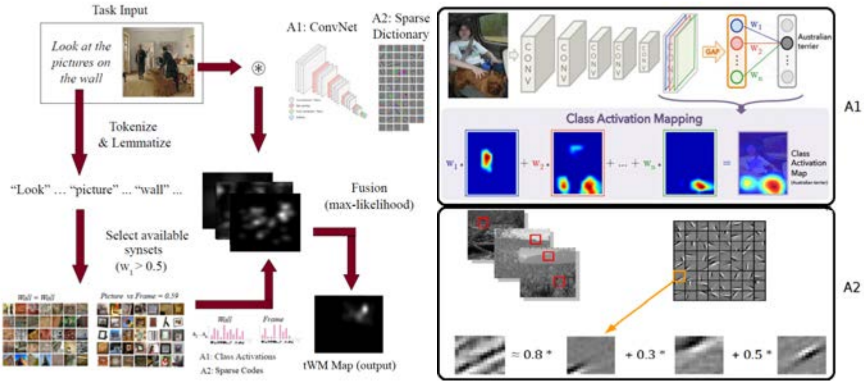


Figure 6.3 – **Left:** Pipeline of the vTE. On the one side, the image is processed by either A1 (CNN architecture) or A2 (sparse dictionary) to obtain a distinct feature maps. On the other side, the sentence is parsed to a specific word embeddings, its semantic similarity with the learned classes will serve to weight the visual representations. **Right:** A1: Illustration of class activation maps at distinct layers of a CNN, from [387]. A2: Schematic representation of sparse coding procedure (obtaining sparse coefficients given a dictionary) from [191]. Each of the feature basis from the dictionary are convolved with the image and weighted (using the coefficients) to obtain class-relevant maps.

6.4 Results on predicting object attention

We have processed PASCAL-S dataset (containing 850 images) and selected the respective 20 annotations from PASCAL VOC 2010. Here we present some examples from sentences "Look for ..." with the respective classes to each annotation. We have calculated the Saliency Index (SI) [23, 297, 300] [Equation 2.18] for comparing the region detection masks and the task working memory maps. Figure 6.4 shows some results for salient object detection given a "Look for..." task for specific categories of objects. Results on SI show that vTE is able to detect objects with better accuracy for the first alternative (A1: Class Activation Maps from VGG16$_{block5\_pool}$), compared to AIM saliency maps and A2 (Sparse Codes from AIM InfoMax dictionary). The presented results are preliminary and STAR-FCT is under implementation process, more details about future results are detailed in Sections 6.5.1-6.5.2.

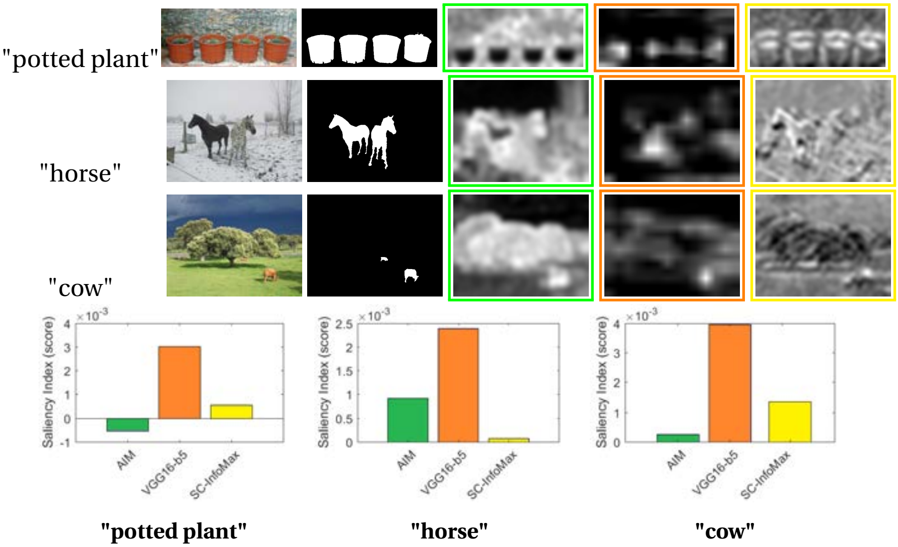


Figure 6.4 – **Top:** Three examples of the task "Look for..." given 3 categories (potted plant, horse and cow) with the corresponding salient region masks (2nd column), AIM saliency maps (3rd column, green), tWM maps with VGG16$_{block5\_pool}$ (4rd column, orange), tWM maps with sparse coding with AIM information maximization dictionary (5th column, yellow). Task Working Memory maps have been obtained from the output of vTE [Figure 6.3-Left]. **Bottom:** Quantitative results of SI for the three category instances.

6.5 Future Work

We plan to include a grammar mechanism to process lexical characteristics of the task sentence (i.e. using imperative english [169, Chapter 2][236][Figure 6.5]), testing as well semantic similarity strategies [199] and word embeddings (e.g. word2vec [203]). Good performance upon complex tasks would include mapping verbs to a vocabulary of actions (localization bias), for both local (specific parts of objects) and global/contextual (on the whole image) guidance [144, Section 4]. Other category weighting strategies could be used by including physiological data from images in the same semantic taxonomy [139]. All these strategies integrated in STAR-FCT can be evaluated by comparing with fixations in image captioning and visual search data.

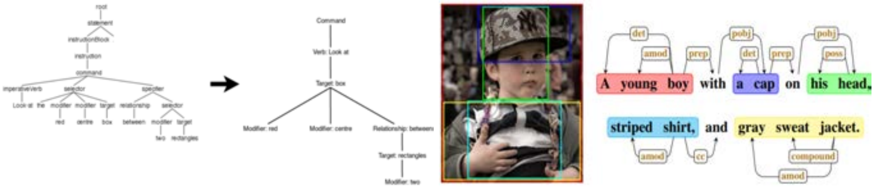


Figure 6.5 – **Left:** Example of how an Imperative English Command (in ANTLR grammar[236]) is able to be parsed to an intermediate representation for the sentence "Look at the red centre box between two rectangles", from [169]. **Right:** Example of image caption and the lexical relationship between words, from [74].

6.5.1 Evaluating task and attention with Image Captions

We plan to evaluate scanpaths from STAR-FC with fixation data, object localization and image captions (e.g. masks and captions from Microsoft COCO and fixations from LSUN17 challenge [135]) or question answering experiments [80], defining a new baseline towards robust computations of visual attention and task priming. We could demonstrate how task can improve detection (compared to no-task) by tuning perception of certain objects, showing how and when are they memorized (accounting for inhibition of return mechanisms) and which visual features are involved.

Attention models [335, 371] based on sequential architectures (i.e. RNN & LSTM [206, 304]) tried to predict image captions by training images with word sequences, by learning the statistical relationship between image representations, words and their sequence. We aim in future work to also evaluate distinct attention strategies and class-wise object attention (e.g. STNet [36], Priming Neural Networks [265]) with our symbolic representations of the sentence, abling to test which CNN arquitectures [6] work best and at which convolutional layers our working memory maps provide better performance in the STAR-FCT architecture.

6.5.2 Evaluating STAR-FCT for Visual Search

A main task to test with new task module could be to assess how good is the model on performing visual search tasks. According to the specification of the feature to be searched in the visual hierarchy, the model would require to efficiently bias the priority of some features over others. In order to do so, we will use a set of psychophysical pattern images (e.g. according to specific color, orientation and spatial frequency). The model would need to, improve the performance (in terms of saccade sequence error with respect to the target centroid and number of fixations inside the search target region) in comparison to the model without task demands. Such procedure could allow to see how well it performs parallel and serial processing of features (for the case of feature and conjunctive search).

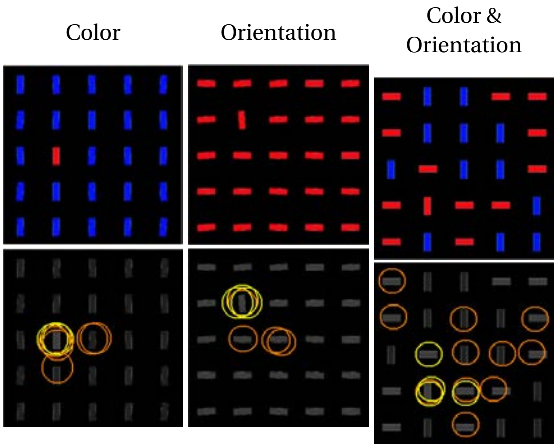


Figure 6.6 – Synthetic image patterns, representing feature (columns 1 and 2) and conjunctive search (column 3). Orange circles depict examples of possible fixations without the task-specific pull and yellow represents the selectively tuned from the task relevance. Adapted from [213].

Another search task to perform is object localization given task demands such as "Look at the handicap sign." or "Where is the handicap sign?". The model would need to parse the input text of the task demands, generate a task graph with its relations and to output the relevance of these categories, finally biasing their priority with respect to other objects in the scene.

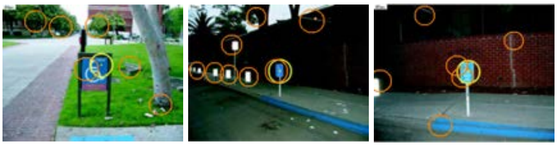
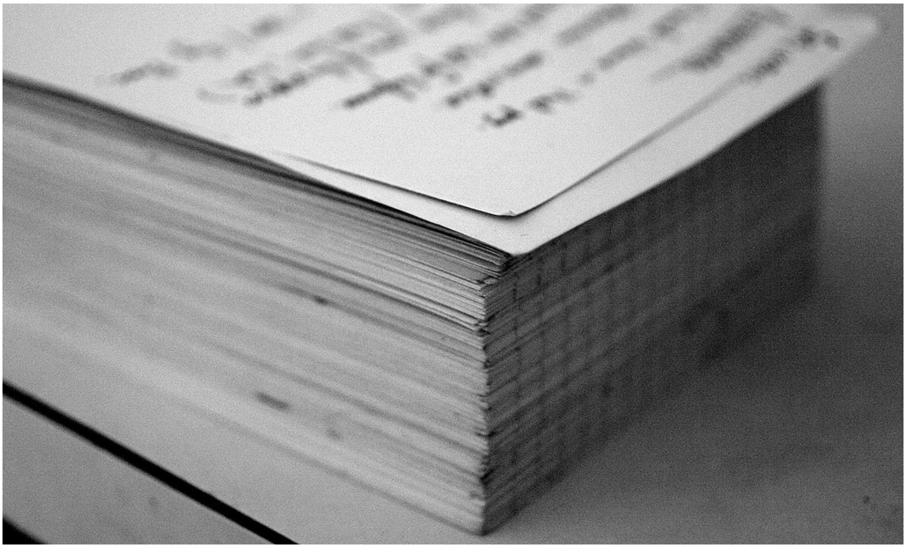


Figure 6.7 – Object localization example for the task "Look for the Handicap Sign". Orange marker represent random locations that appear to be salient. The yellow marker represent where the model should fixate upon the task demands. Adapted from [213].



What did we get until here?

Clausula Part IV

7 Conclusions

7.1 Summary and Contribution

We focus this thesis in understanding visual attention, in concrete, how eyes move while observing scenes and how our brain processes these signals. In this thesis we have proposed to design *biologically plausible* algorithms mimicking attentional mechanisms found in the HVS, that both best represent responses in brain and try to generalize for different scenarios. In this thesis we have conducted experimentation in eye movements with human participants, designed models for saliency prediction, visual search and influences of task. Final conclusions and the list of contributions for each chapter are summarized below:

Chapter 2 - Previous theories explaining how humans attend to visual objects usually test performance (pressing a button upon detection) while searching for visual targets. In our psychophysical experimentation, we explicitly tested this with eye tracking technology using pop-out stimuli (given a salient target and a set of distractors). We parametrized difficulty (as low-level feature contrast) for detecting objects for free-viewing: (1) Corner Angle, (2-3) Visual Segmentation, (4) Contour Integration and (5) Perceptual Grouping; and visual search conditions: (6) Feature and Conjunctive Search, (7) Search Asymmetries, (8) Roughness, (9-10) Color and Brightness contrast, (11) Size contrast, (12) Orientation contrast in (13) Heterogeneous, (14) Nonlinear and (15) Categorical search. Results showed:

1. Eye movements are dependent on feature context (i.e. features that compose the scene)
2. Saliency depends on feature contrast (i.e. differences between features)
3. Fixation patterns vary temporally (i.e. first fixations according to saliency and late ones to relevance)
4. Tasks are able to modulate attention and difficulty of finding salient objects (i.e. in Visual Search is easier than Free-Viewing)

5. Center biases also vary on time and task (i.e. are specially found in late fixations from Free-Viewing)

Chapter 3 Eye movement datasets for saliency comparison tend to use real images and do not parametrize for how salient objects are. We have generated a dataset (SID4VAM) of synthetic images, including fixations in free-viewing and visual search for pop-out psychophysical pattern stimuli. With such dataset, we have labeled each stimuli with a specific difficulty based on feature contrast, abling to observe how saliency models perform on predicting fixations as well as how are they able to locate the salient object. We have also provided detail about temporality of predicted fixations and their relationship with the center bias. From SID4VAM we have computed a benchmark that challenges the state of the art in saliency modeling and provides some concepts that should be accounted when computing saliency metrics:

1. Contrary to current hypotheses, saliency models based on Deep Learning (e.g. SALICON and DeepGazeII) do not perform well on predicting saliency, these are outperformed by ones from other inspiration, specially by Spectral/Fourier inspiration (e.g. HFT and WMAP)
2. Model performance is highly dependent on low-level feature type and contrast
3. Model performance changes upon fixation number (i.e. temporality of fixations)

Chapter 4 - The main basis of our work has been to modelize early visual pathways of the HVS responsible of saliency. In our NSWAM model, we have functionally represented activity from retinal and LGN signals using real images. To do that, we transformed the image to chromatic opponencies transforming image RGB to CIE Lab space and then simple cell responses for each opponency with a self-invertible discrete wavelet transform. These signals have been fed to a network of firing rate neurons simulating lateral/horizontal connections in V1. Through WTA-like mechanisms from these neuronal activity we generate a unique saliency map. This novel approach not only explain saliency computations but also has shown that the same model of V1 is able to simultaneously reproduce (using the same parametrization and without applying any training or optimization procedure) Brightness [239] and Chromatic Induction [65] as well as Visual Discomfort [241]. By testing NSWAM, we can hypothesize the following:

1. Horizontal/lateral connections in V1 could be responsible for the computation of saliency, specially for low-level features

2. NSWAM has a trend to acquire state of the art results for saliency, specially for scenes that lack top-down/contextual influences (i.e. nature and synthetic images)
3. Our model also presents highest performance at highest contrast from feature singleton stimuli (where salient objects pop-out easily) compared with other biologically inspired models

Chapter 5 - We extended the NSWAM saliency model to predict scanpaths (NSWAM-CM) in addition to feedback connections. We been able to provide detail of low-level feature processing during several fixations, and we observed that executive and early visual processes interplay in both bottom-up and top-down attention. A cortical log-polar mapping has been added in order to provide distinct views of the scene, improving results on saliency and providing statistics of saccade sequences. Despite its plausability with brain physiology, cortical magnification mechanisms are not commonly used in the computer vision community. Moreover, we have modeled a simulation of inhibition of return and top-down feedback mechanisms, abling to tune the model to perform better both in free-viewing as well as in feature and categorical search. After testing NSWAM-CM, we can conclude:

1. Cortical Magnification allows to compute distinct views of the image, abling to generate scanpaths and several saliency maps. These computations have shown to improve the previous model predictions.
2. Our model has higher correlation on saccade amplitude compared to other models when predicting visual scanpaths
3. Inhibition of Return mechanisms have shown to improve scanpath predictions, specially when it is parametrized at 1-3 deg of visual angle and a duration of 1-5 times the average fixation time.
4. Top-down feedback can be computed as an inhibitory signal using the same model, abling it to tune the model for feature and categorical search tasks, improving results with respect saliency.

Chapter 6 - Attention towards objects in a scene depend both on on the presented scene and the task to perform. We have extended the Selective Tuning Attentive Reference Fixation Controller model able to predict saccade sequences for a task, defined as a sentence. In that regard, we have proposed several mechanisms able to compute low- and high-level features (using pre-trained CNNs and Sparse Dictionaries) as well as to map the words of the sentence weight visual representations. With these implementations:

1. Deep Learning with CNNs provide better accuracy for generating task working memory maps compared to Sparse Codes of low-level feature dictionaries.
2. Semantic similarity measures are able to determine relationships between categories and representations of features that do not necessarily need to be pre-trained.

We have evaluated current results with salient object localization, and we propose to evaluate the model for predicting fixations in visual search as well as with image captions.

7.2 Future Perspective

This dissertation can been divided in three parts. First, to understand eye movement psychophysics. Second, to model saliency and scanpaths using biologically plausible mechanisms. Third and last, to provide learning methods for tuning attention for visual search and complex task commands. Here we provide some ideas that would extend our work:

Psychophysics: In this work we investigated psychophysical properties of eye movements for low-level feature synthetic image patterns. Current work could be extended by repeating the experimentation for free-viewing conditions (for pop-out stimuli in visual search) in order to understand how task affects low-level feature contrast. Similar psychophysical experimentation could also be done for videos, testing dynamic features such as flicker or motion. Moreover, we could test the influences of pupil size upon search performance given available data (acknowledging that higher pupil size is more related to covert/peripheral attention, whereas lower pupil size is for the case of overt/focal [195]). Another experimentation of interest would be to test the relationship between Gestalt properties (e.g. symmetry [5, 160]) and saliency. This concept could be joined with design principles of "visual weights" and "balance" in art, in addition to aesthetic judgements, abling to test algorithms for predicting best matches in photographic composition.

Saliency and Attention: Even though computational models differ in mechanisms of saliency, we believe pre/post-processing stages can have a big impact on prediction scores [55, 167]. A benchmark testing distinct feature extraction (e.g. DoG, Wavelets, Gabors, Log-Gabors, etc.), fusion (WTA, max-likelihood, inverse, etc.) and normalization (e.g. by energy, by range, etc.) mechanisms could specifically test best low-level saliency algorithms. Further tests with high-level feature computations could be done by generating a dataset with superimposed objects[1] [266] in real scenes (including image captions to provide detail about context and meaning of scenes). This could be set as a baseline to test generalization of current models for several tasks (i.e. in predicting classification, segmentation, captioning, saliency, scanpaths, visual search, etc.).

Multi-task Networks: We plan to extend the NSWAM/NSWAM-CM model by integrating feedforward and feedback mechanisms (simulating intra- and inter-cortical pathways in striate and extrastriate areas) in a multilayer model of the visual cortex. This could be implemented using Spiking Neural Networks [146, 309], which lately showed promising results on classification tasks while preserving biologically-plausible learning mechanisms (i.e. spike-timing dependent plasticity or STDP [194]). Latest transfer learning techniques [231, 330, 374, 378] try to simultaneusly solve problems at distinct domains, but these are not able to retain previous learned features when learning new tasks (b/c catastrophic forgetting) [154]. In that domain, novel Lifelong/Continual learning algorithms [233] need to be proposed to better design architectures able to generalize for distinct tasks simultaneously, something that humans actively perform in our living environments. We would be interested to explore how the brain retains long-term memory formations, as well as how attention, semantic and category embeddings can affect learning and transferability in neural networks.

7.3 Scientific work

7.3.1 Abstracts in National and International Conferences

- David Berga and Xavier Otazu (2016) A Multi-Task Neurodynamical Model of Lateral Interactions in V1: Visual Saliency of Colour Images. 39th European Conference on Visual Perception (ECVP) 2016. [24]

- David Berga and Xavier Otazu (2017) Neurodynamical evidence of gaze prediction decrease with saccade number. 40th European Conference on Visual Perception (ECVP 2017). [25]

- David Berga, Calden Wloka and John K. Tsotsos (2019) Modeling task influences for saccade sequence and visual relevance prediction. 19th Annual Meeting of the Vision Sciences Society (VSS 2019). [32]

[1] Demo from Amir Rosenfeld: https://www.youtube.com/watch?v=qcm3lL4PCC4

- David Berga and Xavier Otazu (2019) Computations of top-down attention by modulating V1 dynamics. MODVIS 2019: Computational and Mathematical Models in Vision. [29]

- David Berga and Xavier Otazu (2019) Computations of inhibition of return mechanisms by modulating V1 dynamics. 28th Annual Computational Neuroscience Meeting (CNS*2019). [27]

- David Berga and Xavier Otazu (2019) Computational modeling of visual attention: What do we know from physiology and psychophysics?. 8th Iberian Conference on Perception (CIP 2019).[28]

- David Berga, Xavier Otazu, Xosé R. Fdez-Vidal, Víctor Leborán and Xosé M. Pardo (2019) Measuring bottom-up visual attention in eye tracking experimentation with synthetic images. 8th Iberian Conference on Perception (CIP 2019). [31]

- David Berga, Xavier Otazu, Xosé R. Fdez-Vidal, Víctor Leborán and Xosé M. Pardo (2019) Generating synthetic images for visual attention modeling. 40th European Conference on Visual Perception (ECVP 2019). [25]

7.3.2 Journal Publications and Conference Proceedings

- David Berga, Xosé R. Fdez-Vidal, Xavier Otazu, Víctor Leborán and Xosé M. Pardo (2019) Psychophysical evaluation of individual low-level feature influences on visual attention. **Vision Research** 154:60-79. [23]

- David Berga and Xavier Otazu (2018) A Neurodynamic model of Saliency prediction in V1. arXiv preprint arXiv:1811.06308. Under Review in **IEEE Transactions on Image Processing.** [26]

- David Berga, Xosé R. Fdez-Vidal, Xavier Otazu and Xosé M. Pardo (2019) SID4VAM: Synthetic Image Dataset for Visual Attention Modeling. Under Review in **International Conference in Computer Vision (ICCV)** 2019.

- David Berga and Xavier Otazu (2019) Modeling Bottom-Up and Top-Down Attention with a Neurodynamic Model of V1. arXiv preprint arXiv:1904.02741. Under Review in **PLOS Computational Biology.** [30]

Bibliography

[1] Ala Aboudib, Vincent Gripon, and Gilles Coppin. A model of bottom-up visual attention using cortical magnification. In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, apr 2015.

[2] Radhakrishna Achanta, Sheila Hemami, Francisco Estrada, and Sabine Susstrunk. Frequency-tuned salient region detection. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2009.

[3] Hossein Adeli, Françoise Vitu, and Gregory J. Zelinsky. A model of the superior colliculus predicts fixation locations during scene viewing and visual search. *The Journal of Neuroscience*, 37(6):1453–1467, dec 2016.

[4] Mehran Ahmadlou, Larry S. Zweifel, and J. Alexander Heimel. Functional modulation of primary visual cortex by the superior colliculus in the mouse. *Nature Communications*, 9(1), sep 2018.

[5] Arash Akbarinia, C. Alejandro Parraga, Marta Expósito, Bogdan Raducanu, and Xavier Otazu. Can biological solutions help computers to detect symmetry? In *40th European Conference on Visual Perception (ECVP)*, pages 95–95, 2017.

[6] Md Zahangir Alom, Tarek M. Taha, Christopher Yakopcic, Stefan Westberg, Paheding Sidike, Mst Shamima Nasrin, Brian C Van Esesn, Abdul A S. Awwal, and Vijayan K. Asari. The history began from alexnet: A comprehensive survey on deep learning approaches, 2018.

[7] Kinjiro Amano and David H. Foster. Influence of local scene color on fixation position in visual search. *Journal of the Optical Society of America A*, 31(4):A254, feb 2014.

[8] James R. Antes. The time course of picture viewing. *Journal of Experimental Psychology*, 103(1):62–70, 1974.

[9] Ján Antolík and James A. Bednar. Development of maps of simple and complex cells in the primary visual cortex. *Frontiers in Computational Neuroscience*, 5, 2011.

[10] Akiyuki Anzai, Xinmiao Peng, and David C Van Essen. Neurons in monkey visual area v2 encode combinations of orientations. *Nature Neuroscience*, 10(10):1313–1321, sep 2007.

[11] Martin A. Asenov. Dynamic model of interactions between orientation selective neurons in primary visual cortex. Master's thesis, University of Edinburg, Edinburgh, UK, 2016.

[12] Aymen Azaza, Joost van de Weijer, Ali Douik, and Marc Masana. Context proposals for saliency detection. *Computer Vision and Image Understanding*, 174:1–11, September 2018.

[13] William F. Bacon and Howard E. Egeth. Overriding stimulus-driven attentional capture. *Perception & Psychophysics*, 55(5):485–496, sep 1994.

[14] Moshe Bar and Irving Biederman. Subliminal visual priming. *Psychological Science*, 9(6):464–468, November 1998.

[15] H. B. Barlow. Redundancy reduction revisited. *Network*, 12 3:241–53, 2001.

[16] David GT Barrett, Ari S Morcos, and Jakob H Macke. Analyzing biological and artificial neural networks: challenges with opportunities for synergy? *Current Opinion in Neurobiology*, 55:55–64, April 2019.

[17] Ben Bauer, Pierre Jolicoeur, and William B Cowan. Distractor heterogeneity versus linear separability in colour visual search. *Perception*, 25(11):1281–1293, nov 1996.

[18] Andrew H. Bell, Tatiana Pasternak, and Leslie G. Ungerleider. *The new visual neurosciences*. The MIT Press, Cambridge, Massachusetts, oct 2013.

[19] Mercedes Barchilon Ben-Av and Dov Sagi. Perceptual grouping by similarity and proximity: Experimental results can be predicted by intensity autocorrelations. *Vision Research*, 35(6):853–866, mar 1995.

[20] Mercedes Barchilon Ben-Av, Dov Sagi, and Jochen Braun. Visual attention and perceptual grouping. *Perception & Psychophysics*, 52(3):277–294, may 1992.

[21] Yoshua Bengio, Dong-Hyun Lee, Jorg Bornschein, Thomas Mesnard, and Zhouhan Lin. Towards biologically plausible deep learning, 2015.

[22] David Berga, Xosé R. Fdez-Vidal, Xavier Otazu, Víctor Leborán, and Xosé M. Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60 – 79, 2019.

[23] David Berga, Xosé R. Fdez-Vidal, Xavier Otazu, Víctor Leborán, and Xosé M. Pardo. Psychophysical evaluation of individual low-level feature influences on visual attention. *Vision Research*, 154:60 – 79, 2019.

[24] David Berga and Xavier Otazu. A multi-task neurodynamical model of lateral interactions in v1: Visual saliency of colour images. In *39th European Conference on Visual Perception (ECVP)*, pages 51–51. SAGE PUBLICATIONS LTD, 2016.

[25] David Berga and Xavier Otazu. Neurodynamical evidence of gaze prediction decrease with saccade number. In *40th European Conference on Visual Perception (ECVP)*, pages 88–88, 2017.

[26] David Berga and Xavier Otazu. A neurodynamic model of saliency prediction in v1. *arXiv preprint arXiv:1811.06308*, 2018.

[27] David Berga and Xavier Otazu. Computations of inhibition of return mechanisms by modulating v1 dynamics. In *28th Annual Computational Neuroscience Meeting (CNS*2019)*, 2019.

[28] David Berga and Xavier Otazu. Computations of inhibition of return mechanisms by modulating v1 dynamics. In *8th Iberian Conference on Perception (CIP 2019). The Spanish Journal of Psychology.*, 2019.

[29] David Berga and Xavier Otazu. Computations of top-down attention by modulating v1 dynamics. In *MODVIS 2019: Computational and Mathematical Models in Vision*, 2019.

[30] David Berga and Xavier Otazu. Modeling bottom-up and top-down attention with a neurodynamic model of v1. April 2019.

[31] David Berga, Xavier Otazu, Xosé R. Fdez-Vidal, Víctor Leborán, and Xosé M. Pardo. Measuring bottom-up visual attention in eye tracking experimentation with synthetic images. In *8th Iberian Conference on Perception (CIP 2019). The Spanish Journal of Psychology.*, 2019.

[32] David Berga, Calden Wloka, and John Tsotsos. Modeling task influences for saccade sequence and visual relevance prediction. *19th Annual Meeting of the Vision Sciences Society (VSS 2019). Journal of Vision.*, 2019.

[33] Matthias Bethge, Matthias Kümmerer, and Thomas Wallis. How close are we to understanding image-based saliency?, 2015.

[34] Rushi Bhatt, Gail A. Carpenter, and Stephen Grossberg. Texture segregation by visual cortex: Perceptual grouping, attention, and learning. *Vision Research*, 47(25):3173–3211, nov 2007.

[35] Irving Biederman, Robert J. Mezzanotte, and Jan C. Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive Psychology*, 14(2):143–177, April 1982.

[36] Mahdi Biparva and John Tsotsos. Stnet: Selective tuning of convolutional networks for object localization. In *The IEEE International Conference on Computer Vision (ICCV) Workshops*, Oct 2017.

[37] P. O. Bishop, J. S. Coombs, and G. H. Henry. Receptive fields of simple cells in the cat striate cortex. *The Journal of Physiology*, 231(1):31–60, May 1973.

[38] James W. Bisley and Michael E. Goldberg. Neural correlates of attention and distractibility in the lateral intraparietal area. *Journal of Neurophysiology*, 95(3):1696–1717, mar 2006.

[39] Giuseppe Boccignone and Mario Ferraro. Modelling gaze shift as a constrained random walk. *Physica A: Statistical Mechanics and its Applications*, 331(1-2):207–218, jan 2004.

[40] A. Borji and L. Itti. Defending yarbus: Eye movements reveal observers' task. *Journal of Vision*, 14(3):29–29, mar 2014.

[41] Ali Borji and Laurent Itti. State-of-the-art in visual attention modeling. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(1):185–207, jan 2013.

[42] Ali Borji and Laurent Itti. Cat2000: A large scale fixation dataset for boosting saliency research. *CVPR 2015 workshop on "Future of Datasets"*, 2015. arXiv preprint arXiv:1505.03581.

[43] Ali Borji, D. N. Sihite, and L. Itti. Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study. *IEEE Transactions on Image Processing*, 22(1):55–69, jan 2013.

[44] Ali Borji, Dicky N. Sihite, and Laurent Itti. What stands out in a scene? a study of human explicit saliency judgment. *Vision Research*, 91:62–77, oct 2013.

[45] Ali Borji and James Tanner. Reconciling saliency and object center-bias hypotheses in explaining free-viewing fixations. *IEEE Transactions on Neural Networks and Learning Systems*, 27(6):1214–1226, jun 2016.

[46] Ali Borji, Hamed R. Tavakoli, Dicky N. Sihite, and Laurent Itti. Analysis of scores, datasets, and models in visual saliency prediction. In *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.

[47] Farran Briggs. *Mammalian Visual System Organization*. Oxford University Press, February 2017.

[48] D. Brockmann. Are human scanpaths levy flights? In *9th International Conference on Artificial Neural Networks: ICANN '99*. IEE, 1999.

[49] N. D. B. Bruce and J. K. Tsotsos. Saliency, attention, and visual search: An information theoretic approach. *Journal of Vision*, 9(3):5–5, mar 2009.

[50] Neil D. B. Bruce and John K. Tsotsos. Saliency based on information maximization. In *Proceedings of the 18th International Conference on Neural Information Processing Systems*, NIPS'05, pages 155–162, Cambridge, MA, USA, 2005. MIT Press.

[51] Neil D.B. Bruce, Calden Wloka, Nick Frosst, Shafin Rahman, and John K. Tsotsos. On computational modeling of visual saliency: Examining what's right, and what's left. *Vision Research*, 116:95–112, nov 2015.

[52] G.T. Buswell. *How People Look at Pictures: A Study of the Psychology of Perception in Art*. University of Chicago Press, 1935.

[53] Z. Bylinskii, E.M. DeGennaro, R. Rajalingham, H. Ruda, J. Zhang, and J.K. Tsotsos. Towards the quantitative evaluation of visual attention models. *Vision Research*, 116:258–268, nov 2015.

[54] Zoya Bylinskii, Tilke Judd, Ali Borji, Laurent Itti, Frédo Durand, Aude Oliva, and Antonio Torralba. Mit saliency benchmark. http://saliency.mit.edu/.

[55] Zoya Bylinskii, Tilke Judd, Aude Oliva, Antonio Torralba, and Fredo Durand. What do different evaluation metrics tell us about saliency models? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 1–1, 2018.

[56] John Byrne. *Learning and memory : a comprehensive reference*. Academic Press is an imprint of Elsevier, Kidlington, Oxford, United Kingdom, 2017.

[57] Giovanni Caputo. Object grouping contingent upon background. *Vision Research*, 37(10):1313–1324, apr 1997.

[58] Matteo Carandini. What simple and complex cells compute. *The Journal of Physiology*, 577(2):463–466, November 2006.

[59] Nancy B. Carlisle and Árni Kristjánsson. How visual working memory contents influence priming of visual attention. *Psychological Research*, 82(5):833–839, April 2017.

[60] Marisa Carrasco. Covert attention increases contrast sensitivity: psychophysical, neurophysiological and neuroimaging studies. In *Visual Perception - Fundamentals of Vision: Low and Mid-Level Processes in Perception*, pages 33–70. Elsevier, 2006.

[61] Marisa Carrasco. Visual attention: The past 25 years. *Vision Research*, 51(13):1484–1525, July 2011.

[62] Marisa Carrasco, Anna Marie Giordano, and Brian McElree. Attention speeds processing across eccentricity: Feature and conjunction searches. *Vision Research*, 46(13):2028–2040, jun 2006.

[63] Marisa Carrasco and Yaffa Yeshurun. The contribution of covert attention to the set-size and eccentricity effects in visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 24(2):673–692, 1998.

[64] M. S. Castelhano, M. L. Mack, and J. M. Henderson. Viewing task influences eye movement control during active scene perception. *Journal of Vision*, 9(3):6–6, mar 2009.

[65] Xim Cerda and Xavier Otazu. A Multi-Task Neurodynamical Model of Lateral Interactions in V1: Chromatic Induction. *39th European Conference of Visual Perception, PERCEPTION*, 45(2):51, 2016.

[66] Hung-Cheng Chang, Stephen Grossberg, and Yongqiang Cao. Where's waldo? how perceptual, cognitive, and emotional brain processes cooperate during learning to categorize and find desired objects in a cluttered scene. *Frontiers in Integrative Neuroscience*, 8, jun 2014.

[67] Sylvain Chevallier, Nicolas Cuperlier, and Philippe Gaussier. Efficient neural models for visual attention. In *Computer Vision and Graphics*, pages 257–264. Springer Berlin Heidelberg, 2010.

[68] Marvin M. Chun and Yuhong Jiang. Contextual cueing: Implicit learning and memory of visual context guides spatial attention. *Cognitive Psychology*, 36(1):28–71, jun 1998.

[69] A. D. F. Clarke, M. J. Chantler, and P. R. Green. Modeling visual search on a rough surface. *Journal of Vision*, 9(4):11–11, apr 2009.

[70] A.D.F. Clarke, P.R. Green, M.J. Chantler, and K. Emrith. Visual search for a target against a 1/f(beta) continuous textured background. *Vision Research*, 48(21):2193–2203, sep 2008.

[71] Alasdair D.F. Clarke and Benjamin W. Tatler. Deriving an appropriate baseline for describing fixation behaviour. *Vision Research*, 102:41–51, sep 2014.

[72] M. Conci, H. J. Muller, and A. von Muhlenen. Object-based implicit learning in visual search: Perceptual segmentation constrains contextual cueing. *Journal of Vision*, 13(3):15–15, jul 2013.

[73] Maurizio Corbetta and Gordon L. Shulman. Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201–215, mar 2002.

[74] Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. Show, control and tell: A framework for generating controllable and grounded captions, 2018.

[75] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. A Deep Multi-Level Network for Saliency Prediction. In *International Conference on Pattern Recognition (ICPR)*, 2016.

[76] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting human eye fixations via an lstm-based saliency attentive model, 2016.

[77] Marcella Cornia, Lorenzo Baraldi, Giuseppe Serra, and Rita Cucchiara. Predicting Human Eye Fixations via an LSTM-based Saliency Attentive Model. *IEEE Transactions on Image Processing*, 27(10):5142–5154, 2018.

[78] S. C. Dakin and N. J. Baruch. Context influences contour integration. *Journal of Vision*, 9(2):13–13, feb 2009.

[79] M. V. Danilova and J. D. Mollon. Symmetries and asymmetries in chromatic discrimination. *Journal of the Optical Society of America A*, 31(4):A247, feb 2014.

[80] Abhishek Das, Harsh Agrawal, C. Lawrence Zitnick, Devi Parikh, and Dhruv Batra. Human attention in visual question answering: Do humans and deep networks look at the same regions?, 2016.

[81] Floor de Groot, Falk Huettig, and Christian N. L. Olivers. When meaning matters: The temporal dynamics of semantic influences on visual attention. *Journal of Experimental Psychology: Human Perception and Performance*, 42(2):180–196, 2016.

[82] Gustavo Deco and Edmund T. Rolls. A neurodynamical cortical model of visual attention and invariant object recognition. *Vision Research*, 44(6):621–642, mar 2004.

[83] R. Desimone. Visual attention mediated by biased competition in extrastriate visual cortex. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 353(1373):1245–1255, aug 1998.

[84] Robert Desimone and John Duncan. Neural mechanisms of selective visual attention. *Annual Review of Neuroscience*, 18(1):193–222, mar 1995.

[85] Peter U. Diehl and Matthew Cook. Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in Computational Neuroscience*, 9, aug 2015.

[86] John Duncan and Glyn W. Humphreys. Visual search and stimulus similarity. *Psychological Review*, 96(3):433–458, 1989.

[87] Michael D'Zmura. Color in visual search. *Vision Research*, 31(6):951–966, jan 1991.

[88] Howard E. Egeth and Steven Yantis. VISUAL ATTENTION: Control, representation, and time course. *Annual Review of Psychology*, 48(1):269–297, feb 1997.

[89] Michelle L. Eisenberg and Jeffrey M. Zacks. Ambient and focal visual processing of naturalistic activity. *Journal of Vision*, 16(2):5, mar 2016.

[90] GÖSta Ekman. Weber's law and related functions. *The Journal of Psychology*, 47(2):343–352, April 1959.

[91] Mazyar Fallah and John H. Reynolds. Contrast dependence of smooth pursuit eye movements following a saccade to superimposed targets. *PLoS ONE*, 7(5):e37888, may 2012.

[92] G. T. Fechner. *Elements of Psychophysics, Volume 1*. Holt, Rinehart and Winston, the University of Michigan, 1966.

[93] J. Fecteau and D. Munoz. Salience, relevance, and firing: a priority map for target selection. *Trends in Cognitive Sciences*, 10(8):382–390, aug 2006.

[94] David J. Field, Anthony Hayes, and Robert F. Hess. Contour integration by the human visual system: Evidence for a local "association field". *Vision Research*, 33(2):173–193, jan 1993.

[95] Sylvain Fischer, Filip Šroubek, Laurent Perrinet, Rafael Redondo, and Gabriel Cristóbal. Self-invertible 2d log-gabor wavelets. *International Journal of Computer Vision*, 75(2):231–246, jan 2007.

[96] Alban Flachot and Karl R. Gegenfurtner. Processing of chromatic information in a deep convolutional neural network. *Journal of the Optical Society of America A*, 35(4):B334, March 2018.

[97] Jonathan Flombaum. Visual search for features and conjunctions. *J. Vis. Exp.*, 2015.

[98] Brice Follet, Olivier Le Meur, and Thierry Baccino. New insights into ambient and focal visual fixations using an automatic classification algorithm. *i-Perception*, 2(6):592–610, 2011.

[99] Simone Frintrop, Erich Rome, and Henrik I. Christensen. Computational visual attention systems and their cognitive foundations. *ACM Transactions on Applied Perception*, 7(1):1–39, January 2010.

[100] Kunihiko Fukushima. Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition unaffected by shift in position. *Biological Cybernetics*, 36(4):193–202, April 1980.

[101] Dashan Gao. *A discriminant hypothesis for visual saliency: computational principles, biological plausibility and applications in computer vision.* PhD thesis, UC San Diego, 2008.

[102] Antón Garcia-Diaz, Xosé R. Fdez-Vidal, Xosé M. Pardo, and Raquel Dosil. Saliency from hierarchical adaptation through decorrelation and variance normalization. *Image and Vision Computing*, 30(1):51–64, jan 2012.

[103] Isabel Gauthier. Visual priming: The ups and downs of familiarity. *Current Biology*, 10(20):R753–R756, October 2000.

[104] Robert Geirhos, David H. J. Janssen, Heiko H. Schütt, Jonas Rauber, Matthias Bethge, and Felix A. Wichmann. Comparing deep neural networks against humans: object recognition when the signal gets weaker, 2017.

[105] Charles D. Gilbert. Horizontal integration and cortical dynamics. *Neuron*, 9(1):1–13, jul 1992.

[106] Richard Godijn and Jan Theeuwes. Oculomotor capture and inhibition of return: Evidence for an oculomotor suppression account of IOR. *Psychological Research*, 66(4):234–246, nov 2002.

[107] Stas Goferman, Lihi Zelnik-Manor, and Ayellet Tal. Context-aware saliency detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(10):1915–1926, oct 2012.

[108] E. Bruce Goldstein, editor. *Blackwell Handbook of Sensation and Perception*. Blackwell Publishing Ltd, January 2005.

[109] M. González-Audícana, X. Otazu, O. Fors, and A. Seco. Comparison between mallat's and the 'à trous' discrete wavelet transform based algorithms for the fusion of multispectral and panchromatic images. *International Journal of Remote Sensing*, 26(3):595–614, feb 2005.

[110] Melvyn A. Goodale and A.David Milner. Separate visual pathways for perception and action. *Trends in Neurosciences*, 15(1):20–25, January 1992.

[111] Ian Goodfellow. *Deep learning*. The MIT Press, Cambridge, Massachusetts, 2016.

[112] Paula Goolkasian. Size scaling and spatial factors in visual attention. *The American Journal of Psychology*, 110(3):397, 1997.

[113] Michelle R. Greene, Tommy Liu, and Jeremy M. Wolfe. Reconsidering yarbus: A failure to predict observers' task from eye movement patterns. *Vision Research*, 62:1–8, jun 2012.

[114] Stephen Grossberg, Jesse Palma, and Massimiliano Versace. Resonant cholinergic dynamics in cognitive and motor decision-making: Attention, category learning, and choice in neocortex, superior colliculus, and optic tectum. *Frontiers in Neuroscience*, 9, jan 2016.

[115] Yuqiao Gu and Hans Liljenström. A neural network model of attention-modulated neurodynamics. *Cognitive Neurodynamics*, 1(4):275–285, oct 2007.

[116] Chenlei Guo, Qi Ma, and Liming Zhang. Spatio-temporal saliency detection using phase spectrum of quaternion fourier transform. In *2008 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2008.

[117] Jr. H. L. Pick. *Modes of Perceiving and Processing Information*. Psychology Press, March 2014.

[118] Isma Hadji and Richard P. Wildes. What do we understand about convolutional networks?, 2018.

[119] Jonathan Harel, Christof Koch, and Pietro Perona. Graph-based visual saliency. In B. Schölkopf, J. C. Platt, and T. Hoffman, editors, *Advances in Neural Information Processing Systems 19*, pages 545–552. MIT Press, 2007.

[120] Bernard M. Hart, Hannah C. E. F. Schmidt, Christine Roth, and Wolfgang Einhäuser. Fixations on objects in natural scenes: dissociating importance from salience. *Frontiers in Psychology*, 4, 2013.

[121] William G. Hayward and Darren Burke. What pops-out in pop-out?. In *41st Annual Meeting of the Psychonomic Society*. IEEE, 2000.

[122] David J. Heeger. Normalization of cell responses in cat striate cortex. *Visual Neuroscience*, 9(02):181–197, August 1992.

[123] John M. Henderson, Jr. Weeks Phillip A., and Andrew Hollingworth. The effects of semantic consistency on eye movements during complex scene viewing. *Journal of Experimental Psychology: Human Perception and Performance*, 25(1):210–228, 1999.

[124] Jeanny Herault. *Biologically inspired computer vision : fundamentals and applications*. Wiley, Hoboken, N.J, 2015.

[125] R.F Hess, A Hayes, and D.J Field. Contour integration and cortical processing. *Journal of Physiology-Paris*, 97(2-3):105–119, mar 2003.

[126] Shaul Hochstein and Merav Ahissar. View from the top. *Neuron*, 36(5):791–804, December 2002.

[127] Andrew Hollingworth. Does consistent scene context facilitate object perception? *Journal of Experimental Psychology: General*, 127(4):398–415, 1998.

[128] Anja K.E. Horn and Christopher Adamczyk. Reticular formation. In *The Human Nervous System*, pages 328–366. Elsevier, 2012.

[129] Todd S. Horowitz and Jeremy M. Wolfe. Visual search has no memory. *Nature*, 394(6693):575–577, aug 1998.

[130] Xiaodi Hou, J. Harel, and C. Koch. Image signature: Highlighting sparse salient regions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 34(1):194–201, jan 2012.

[131] Xiaodi Hou and Liqing Zhang. Saliency detection: A spectral residual approach. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2007.

[132] Xiaodi Hou and Liqing Zhang. Dynamic visual attention: searching for coding length increments. In D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, editors, *Advances in Neural Information Processing Systems 21*, pages 681–688. Curran Associates, Inc., 2009.

[133] Kesong Hu, Junya Zhan, Bingzhao Li, Shuchang He, and Arthur G. Samuel. Multiple cueing dissociates location- and feature-based repetition effects. *Vision Research*, 101:73–81, aug 2014.

[134] Liqiang Huang and Harold Pashler. A boolean map theory of visual attention. *Psychological Review*, 114(3):599–631, 2007.

[135] Xun Huang, Chengyao Shen, Xavier Boix, and Qi Zhao. SALICON: Reducing the semantic gap in saliency prediction by adapting deep neural networks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, dec 2015.

[136] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurones in the cat's striate cortex. *The Journal of Physiology*, 148(3):574–591, October 1959.

[137] D. H. Hubel and T. N. Wiesel. Receptive fields and functional architecture of monkey striate cortex. *The Journal of Physiology*, 195(1):215–243, March 1968.

[138] Alexander C. Huk and David J. Heeger. Task-related modulation of visual cortex. *Journal of Neurophysiology*, 83(6):3525–3536, jun 2000.

[139] Alexander G. Huth, Shinji Nishimoto, An T. Vu, and Jack L. Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, 76(6):1210–1224, December 2012.

[140] Alex D. Hwang, Hsueh-Cheng Wang, and Marc Pomplun. Semantic guidance of eye movements in real-world scenes. *Vision Research*, 51(10):1192–1205, may 2011.

[141] Yasushi Imamoto and Yoshinori Shichida. Cone visual pigments. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1837(5):664–673, may 2014.

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

[143] Laurent Itti. Visual salience. *Scholarpedia*, 2(9):3327, 2007.

[144] Laurent Itti and Ali Borji. *Computational Models: Bottom-Up and Top-Down Aspects.* Oxford University Press, April 2014.

[145] Laurent Itti and Christof Koch. A saliency-based search mechanism for overt and covert shifts of visual attention. *Vision Research*, 40(10-12):1489–1506, jun 2000.

[146] E.M. Izhikevich. Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15(5):1063–1070, sep 2004.

[147] Ming Jiang, Shengsheng Huang, Juanyong Duan, and Qi Zhao. SALICON: Saliency in context. In *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2015.

[148] E. N. Johnson, M. J. Hawken, and R. Shapley. The orientation selectivity of color-responsive neurons in macaque v1. *Journal of Neuroscience*, 28(32):8096–8106, aug 2008.

[149] John Jonides. *Voluntary versus automatic control over the mind's eye's movement Attention and performance IX.* Erlbaum Associates, 1981.

[150] Tilke Judd, Fredo Durant, and Antonio Torralba. A benchmark of computational models of saliency to predict human fixations. *CSAIL Technical Reports*, jan 2012.

[151] Tilke Judd, Krista Ehinger, Fredo Durand, and Antonio Torralba. Learning to predict where humans look. In *2009 IEEE 12th International Conference on Computer Vision*. IEEE, sep 2009.

[152] Eric Kandel. *Principles of neural science.* McGraw-Hill, Health Professions Division, New York, 2000.

[153] Fahad Shahbaz Khan, Joost van de Weijer, and Maria Vanrell. Modulating shape features by color attention for object recognition. *International Journal of Computer Vision*, 98(1):49–64, September 2011.

[154] James Kirkpatrick, Razvan Pascanu, Neil Rabinowitz, Joel Veness, Guillaume Desjardins, Andrei A. Rusu, Kieran Milan, John Quan, Tiago Ramalho, Agnieszka Grabska-Barwinska, Demis Hassabis, Claudia Clopath, Dharshan Kumaran, and Raia Hadsell. Overcoming catastrophic forgetting in neural networks, 2016.

[155] Werner Klotz and Odmar Neumann. Motor activation without conscious discrimination in metacontrast masking. *Journal of Experimental Psychology: Human Perception and Performance*, 25(4):976–992, 1999.

[156] E. Kobatake and K. Tanaka. Neuronal selectivities to complex object features in the ventral visual pathway of the macaque cerebral cortex. *Journal of Neurophysiology*, 71(3):856–867, March 1994.

[157] Christof Koch and Shimon Ullman. Shifts in selective visual attention: Towards the underlying neural circuitry. In *Matters of Intelligence*, pages 115–141. Springer Netherlands, 1987.

[158] Kurt Koffka. *Principles of gestalt psychology*. Harbrace J, US, 1935.

[159] Garry Kong, David Alais, and Erik Van der Burg. Orientation categories used in guidance of attention in visual search can differ in strength. *Attention, Perception, & Psychophysics*, jul 2017.

[160] Gert Kootstra, Bart de Boer, and Lambert R. B. Schomaker. Predicting eye fixations on complex visual stimuli using local symmetry. *Cognitive Computation*, 3(1):223–240, jan 2011.

[161] Eileen Kowler. Eye movements: The past 25years. *Vision Research*, 51(13):1457–1483, jul 2011.

[162] Árni Kristjánsson and Jon Driver. Priming in visual search: Separating the effects of target repetition, distractor repetition and role-reversal. *Vision Research*, 48(10):1217–1232, May 2008.

[163] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, 60(6):84–90, May 2017.

[164] N. Kruger, P. Janssen, S. Kalkan, M. Lappe, A. Leonardis, J. Piater, A. J. Rodriguez-Sanchez, and L. Wiskott. Deep hierarchies in the primate visual cortex: What can we learn for computer vision? *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(8):1847–1871, aug 2013.

[165] Hideyuki Kubota, Yusuke Sugano, Takahiro Okabe, Yoichi Sato, Akihiro Sugimoto, and Kazuo Hiraki. Incorporating visual field characteristics into a saliency map. In *Proceedings of the Symposium on Eye Tracking Research and Applications - ETRA '12*. ACM Press, 2012.

[166] Michael Kubovy and James R. Pomerantz, editors. *Perceptual Organization*. Routledge, March 2017.

[167] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Saliency benchmarking made easy: Separating models, maps and metrics. In *Computer Vision – ECCV 2018*, pages 798–814. Springer International Publishing, 2018.

[168] Matthias Kummerer, Thomas S.A. Wallis, Leon A. Gatys, and Matthias Bethge. Understanding low- and high-level contributions to fixation prediction. In *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, oct 2017.

[169] Toni Kunic. Cognitive program compiler. Master's thesis, York University, Toronto, Ontario, Canada, 2017.

[170] Phillipp Kurtz, Katharine A. Shapcott, Jochen Kaiser, Joscha T. Schmiedt, and Michael C. Schmid. The influence of endogenous and exogenous spatial attention on decision confidence. *Scientific Reports*, 7(1), jul 2017.

[171] Matthias Kümmerer, Thomas S. A. Wallis, and Matthias Bethge. Deepgaze ii: Reading fixations from deep features trained on object recognition, 2016.

[172] Victor A.F. Lamme and Pieter R. Roelfsema. The distinct modes of vision offered by feedforward and recurrent processing. *Trends in Neurosciences*, 23(11):571–579, nov 2000.

[173] Otto Lappi. Eye movements in the wild: Oculomotor control, gaze behavior & frames of reference. *Neuroscience & Biobehavioral Reviews*, 69:49–68, October 2016.

[174] An T.D. Le, Jasmine Payne, Charlotte Clarke, Murphy A. Kelly, Francesca Prudenziati, Elise Armsby, Olivier Penacchio, and Arnold J. Wilkins. Discomfort from urban scenes: Metabolic consequences. *Landscape and Urban Planning*, 160:61–68, apr 2017.

[175] Victor Leboran, Anton Garcia-Diaz, Xose R. Fdez-Vidal, and Xose M. Pardo. Dynamic whitening saliency. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39(5):893–907, may 2017.

[176] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, May 2015.

[177] JOSEF LEDERER. BRAINEDNESS, HANDEDNESS AND EYEDNESS: THE MEANING OF OCULAR DOMINANCE. *Clinical and Experimental Optometry*, 53(11):323–347, nov 1970.

[178] Jung H. Lee, Christof Koch, and Stefan Mihalas. A computational analysis of the function of three inhibitory cell types in contextual visual processing. *Frontiers in Computational Neuroscience*, 11, apr 2017.

[179] Tai Sing Lee. Image representation using 2d gabor wavelets. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 18(10):959–971, 1996.

[180] Olivier LeMeur and Thierry Baccino. Methods for comparing scanpaths and saliency maps: strengths and weaknesses. *Behavior Research Methods*, 45(1):251–266, jul 2012.

[181] P Lennie, J Krauskopf, and G Sclar. Chromatic mechanisms in striate cortex of macaque. *The Journal of Neuroscience*, 10(2):649–669, feb 1990.

[182] Jian Li, Martin D. Levine, Xiangjing An, Xin Xu, and Hangen He. Visual saliency based on scale-space analysis in the frequency domain. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(4):996–1010, apr 2013.

[183] Wu Li and Charles D. Gilbert. Global contour saliency and local colinear interactions. *Journal of Neurophysiology*, 88(5):2846–2856, nov 2002.

[184] Zhaoping Li. A neural model of contour integration in the primary visual cortex. *Neural Computation*, 10(4):903–940, may 1998.

[185] Zhaoping Li. Visual segmentation by contextual influences via intra-cortical interactions in the primary visual cortex. *Network: Computation in Neural Systems*, 10(2):187–212, 1999. PMID: 10378191.

[186] Zhaoping Li. Pre-attentive segmentation in the primary visual cortex. *Spatial Vision*, 13(1):25–50, jan 2000.

[187] Tony Lindeberg. A computational theory of visual receptive fields. *Biological Cybernetics*, 107(6):589–635, November 2013.

[188] Fernando Lopez-Garcia, Xose Ramon, Xose Manuel, and Raquel Dosil. Scene recognition through visual attention and image features: A comparison between SIFT and SURF approaches. In *Object Recognition*. InTech, apr 2011.

[189] Donald I. A. MacLeod and Robert M. Boynton. Chromaticity diagram showing cone excitation by stimuli of equal luminance. *Journal of the Optical Society of America*, 69(8):1183, aug 1979.

[190] Lamberto Maffei and Adriana Fiorentini. The visual cortex as a spatial frequency analyser. *Vision Research*, 13(7):1255–1267, jul 1973.

[191] Julien Mairal, Michael Elad, and Guillermo Sapiro. Sparse representation for color image restoration. *IEEE Transactions on Image Processing*, 17(1):53–69, January 2008.

[192] S.K. Mannan, D.S. Wooding, and K.H. Ruddock. The relationship between the locations of spatial features and those of fixations made during visual examination of briefly presented images. *Spatial Vision*, 10(3):165–188, jan 1996.

[193] Mateja Marić and Dražen Domijan. A neurodynamic model of feature-based spatial selection. *Frontiers in Psychology*, 9, mar 2018.

[194] Timothée Masquelier and Simon J. Thorpe. Unsupervised learning of visual features through spike timing dependent plasticity. *PLoS Computational Biology*, 3(2):e31, 2007.

[195] Sebastiaan Mathôt, Lotje van der Linden, Jonathan Grainger, and Françoise Vitu. The pupillary light response reveals the focus of covert visual attention. *PLoS ONE*, 8(10):e78168, October 2013.

[196] K. C. McDermott, G. Malkoc, J. B. Mulligan, and M. A. Webster. Adaptation and visual salience. *Journal of Vision*, 10(13):17–17, nov 2010.

[197] N. V. Kartheek Medathati, Heiko Neumann, Guillaume S. Masson, and Pierre Kornprobst. Bio-inspired computer vision: Towards a synergistic approach of artificial and biological vision. *Computer Vision and Image Understanding*, 150:1–30, September 2016.

[198] David A. Mély and Thomas Serre. Towards a theory of computation in the visual cortex. In *Computational and Cognitive Neuroscience of Vision*, pages 59–84. Springer Singapore, oct 2016.

[199] Lingling Meng, Runqing Huang, and Junzhong Gu. A review of semantic similarity measures in wordnet. *International Journal of Hybrid Information Technology*, 6(1):1–12, 2013.

[200] W H Merigan and J H R Maunsell. How parallel are the primate visual pathways? *Annual Review of Neuroscience*, 16(1):369–402, March 1993.

[201] Olivier Le Meur, Antoine Coutrot, Zhi Liu, Adrien Le Roch, Andrea Helo, and Pia Rama. Computational model for predicting visual fixations from childhood to adulthood. *CoRR*, abs/1702.04657, 2017.

[202] Olivier Le Meur and Zhi Liu. Saccadic model of eye movements for free-viewing condition. *Vision Research*, 116:152–164, nov 2015.

[203] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space, 2013.

[204] George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to WordNet: An On-line Lexical Database*. *International Journal of Lexicography*, 3(4):235–244, 12 1990.

[205] Kenichiro Miura, Kazuyo Suehiro, Miyuki Yamamoto, Yasushi Kodaka, and Kenji Kawano. Initiation of smooth pursuit in humans. *Experimental Brain Research*, 141(2):242–249, nov 2001.

[206] Volodymyr Mnih, Nicolas Heess, Alex Graves, and Koray Kavukcuoglu. Recurrent models of visual attention, 2014.

[207] Patrick Monnier and Steven K. Shevell. Chromatic induction from s-cone patterns. *Vision Research*, 44(9):849–856, apr 2004.

[208] N. Murray, M. Vanrell, X. Otazu, and C. A. Parraga. Low-level spatiochromatic grouping for saliency estimation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(11):2810–2816, nov 2013.

[209] Naila Murray, Maria Vanrell, Xavier Otazu, and C. Alejandro Parraga. Saliency estimation using a non-parametric low-level vision model. In *CVPR 2011*. IEEE, jun 2011.

[210] Allen L. Nagy. Interactions between achromatic and chromatic mechanisms in visual search. *Vision Research*, 39(19):3253–3266, oct 1999.

[211] Ken Nakayama and Gerald H. Silverman. Serial and parallel processing of visual feature conjunctions. *Nature*, 320(6059):264–265, mar 1986.

[212] Jonathan J. Nassi and Edward M. Callaway. Parallel processing strategies of the primate visual system. *Nature Reviews Neuroscience*, 10(5):360–372, apr 2009.

[213] Vidhya Navalpakkam and Laurent Itti. Modeling the influence of task on attention. *Vision Research*, 45(2):205–231, jan 2005.

[214] Hoang L. Nhan and Edward M. Callaway. Morphology of superior colliculus- and middle temporal area-projecting neurons in primate primary visual cortex. *The Journal of Comparative Neurology*, 520(1):52–80, Nov 2011.

[215] Eilen Nordlie, Marc-Oliver Gewaltig, and Hans Ekkehard Plesser. Towards reproducible descriptions of neuronal network models. *PLoS Computational Biology*, 5(8):e1000456, aug 2009.

[216] Hans-Christoph Nothdurft. The conspicuousness of orientation and motion contrast. *Spatial Vision*, 7(4):341–363, jan 1993.

[217] Hans-Christoph Nothdurft. The role of features in preattentive vision: Comparison of orientation, motion and color cues. *Vision Research*, 33(14):1937–1958, sep 1993.

[218] Hans-Christoph Nothdurft. Salience from feature contrast: additivity across dimensions. *Vision Research*, 40(10-12):1183–1201, jun 2000.

[219] Hans-Christoph Nothdurft. Salience and target selection in visual search. *Visual Cognition*, 14(4-8):514–542, aug 2006.

[220] Hans-Christoph Nothdurft. Salience-controlled visual search: Are the brightest and the least bright targets found by different processes? *Visual Cognition*, 13(6):700–732, apr 2006.

[221] H.C. Nothdurft. Sensitivity for structure gradient in texture discrimination tasks. *Vision Research*, 25(12):1957–1968, jan 1985.

[222] H.C. Nothdurft. Texture segmentation and pop-out from orientation contrast. *Vision Research*, 31(6):1073–1078, jan 1991.

[223] Antje Nuthmann, Wolfgang Einhäuser, and Immo Schütz. How well can saliency models predict fixation selection in scenes beyond central bias? a new approach to model evaluation using generalized linear mixed models. *Frontiers in Human Neuroscience*, 11, oct 2017.

[224] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, December 2007.

[225] Bruno A. Olshausen and David J. Field. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature*, 381(6583):607–609, jun 1996.

[226] Stefano Padilla, Ondrej Drbohlav, Patrick R. Green, Andy Spence, and Mike J. Chantler. Perceived roughness of $1/f^{\beta}$ noise surfaces. *Vision Research*, 48(17):1791–1797, aug 2008.

[227] John Palmer. Attention in visual search: Distinguishing four causes of a set-size effect. *Current Directions in Psychological Science*, 4(4):118–123, aug 1995.

[228] John Palmer, Preeti Verghese, and Misha Pavel. The psychophysics of visual search. *Vision Research*, 40(10-12):1227–1268, jun 2000.

[229] Stephen E. Palmer. The effects of contextual scenes on the identification of objects. *Memory & Cognition*, 3(5):519–526, sep 1975.

[230] Junting Pan, Cristian Canton, Kevin McGuinness, Noel E. O'Connor, Jordi Torres, Elisa Sayrol, and Xavier and Giro-i Nieto. Salgan: Visual saliency prediction with generative adversarial networks. In *arXiv*, January 2017.

[231] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, October 2010.

[232] Sebastian Pannasch, Jens R. Helmert, Katharina Roth, Henrik Walter, and Ann-Katrin Herbold. Visual fixation durations and saccade amplitudes: Shifting relationship in a variety of conditions. *Journal of Eye Movement Research*, 2(2), 2008.

[233] German I. Parisi, Ronald Kemker, Jose L. Part, Christopher Kanan, and Stefan Wermter. Continual lifelong learning with neural networks: A review. *Neural Networks*, 113:54–71, May 2019.

[234] Derrick Parkhurst, Klinton Law, and Ernst Niebur. Modeling the role of salience in the allocation of overt visual attention. *Vision Research*, 42(1):107–123, jan 2002.

[235] Derrick Parkhurst and Ernst Niebur. Scene content selected by active vision. *Spatial Vision*, 16(2):125–154, jun 2003.

[236] Terence Parr. *The definitive ANTLR 4 reference*. The Pragmatic Programmers, Frisco, TX, 2014.

[237] Harold Pashler. Familiarity and visual change detection. *Perception & Psychophysics*, 44(4):369–378, jul 1988.

[238] Harold Pashler, Karen Dobkins, and Liqiang Huang. Is contrast just another feature for visual selective attention? *Vision Research*, 44(12):1403–1410, jun 2004.

[239] Olivier Penacchio, Xavier Otazu, and Laura Dempere-Marco. A neurodynamical model of brightness induction in v1. *PLoS ONE*, 8(5):e64086, may 2013.

[240] Olivier Penacchio and Arnold J. Wilkins. Visual discomfort and the spatial distribution of fourier energy. *Vision Research*, 108:1–7, mar 2015.

[241] Olivier Penacchio, Arnold J. Wilkins, Xavier Otazu, and Julie M. Harris. Inhibitory function and its contribution to cortical hyperexcitability and visual discomfort as assessed by a computation model of cortical function. *39th European Conference of Visual Perception, PERCEPTION*, 45(2):51, 2016.

[242] Franco Pestilli, Gerardo Viera, and Marisa Carrasco. How do attention and adaptation affect contrast sensitivity? *Journal of Vision*, 7(7):9, may 2007.

[243] C. Pierrot-Deseilligny, R.M. Müri, C.J. Ploner, B. Gaymard, and S. Rivaud-Péchoux. Cortical control of ocular saccades in humans: a model for motricity. In *Progress in Brain Research*, pages 3–17. Elsevier, 2003.

[244] C. Pierrot-Deseilligny, S. Rivaud, B. Gaymard, and Y. Agid. Cortical control of memory-guided saccades in man. *Experimental Brain Research*, 83(3), feb 1991.

[245] Charles Pierrot-Deseilligny, Dan Milea, and René Müri. Eye movement control by the cerebral cortex. *Current Opinion in Neurology*, 17(1):17–25, feb 2004.

[246] Michael I. Posner. Orienting of attention. *Quarterly Journal of Experimental Psychology*, 32(1):3–25, feb 1980.

[247] Claudio M. Privitera, Thom Carney, Stanley Klein, and Mario Aguilar. Analysis of microsaccades and pupil dilation reveals a common decisional origin during visual search. *Vision Research*, 95:43–50, feb 2014.

[248] Michael J. Proulx. Size matters: Large objects capture attention in visual search. *PLoS ONE*, 5(12):e15293, dec 2010.

[249] Maria Solé Puig, Laura Pérez Zapata, J. Antonio Aznar-Casanova, and Hans Supèr. A role of eye vergence in covert attention. *PLoS ONE*, 8(1):e52955, jan 2013.

[250] Dale Purves. *Neuroscience*. Sinauer Associates, Sunderland, Mass, 2001.

[251] Philip T. Quinlan. Visual feature integration theory: Past, present, and future. *Psychological Bulletin*, 129(5):643–673, 2003.

[252] Subramanian Ramanathan, Harish Katti, Nicu Sebe, Mohan Kankanhalli, and Tat-Seng Chua. An eye fixation database for saliency detection in images. In *Computer Vision – ECCV 2010*, pages 30–43. Springer Berlin Heidelberg, 2010.

[253] Thomas Zoëga Ramsøy and Morten Overgaard. Introspection and subliminal perception. *Phenomenology and the Cognitive Sciences*, 3(1):1–23, 2004.

[254] Ronald A. Rensink and James T. Enns. Preemption effects in visual search: Evidence for low-level grouping. *Psychological Review*, 102(1):101–130, 1995.

[255] Ronald A. Rensink and James T. Enns. Early completion of occluded objects. *Vision Research*, 38(15-16):2489–2505, aug 1998.

[256] Nicolas Riche, Matthieu Duvinage, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Saliency and human fixations: State-of-the-art and study of comparison metrics. In *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.

[257] Nicolas Riche and Matei Mancas. Bottom-up saliency models for still images: A practical review. In *From Human Attention to Computational Attention*, pages 141–175. Springer New York, 2016.

[258] Nicolas Riche and Matei Mancas. Bottom-up saliency models for videos: A practical review. In *From Human Attention to Computational Attention*, pages 177–190. Springer New York, 2016.

[259] Nicolas Riche, Matei Mancas, Bernard Gosselin, and Thierry Dutoit. Rare: A new bottom-up saliency model. In *2012 19th IEEE International Conference on Image Processing*. IEEE, sep 2012.

[260] Micah Richert, Dimitry Fisher, Filip Piekniewski, Eugene M. Izhikevich, and Todd L. Hylton. Fundamental principles of cortical computation: unsupervised learning with prediction, compression and feedback, 2016.

[261] Reuben Rideaux, David R. Badcock, Alan Johnston, and Mark Edwards. Temporal synchrony is an effective cue for grouping and segmentation in the absence of form cues. *Journal of Vision*, 16(11):23, sep 2016.

[262] Maximilian Riesenhuber and Tomaso Poggio. Hierarchical models of object recognition in cortex. *Nature Neuroscience*, 2(11):1019–1025, November 1999.

[263] R.W. Rodieck. Quantitative analysis of cat retinal ganglion cell response to visual stimuli. *Vision Research*, 5(12):583–601, dec 1965.

[264] Edmund Rolls. *Memory, attention, and decision-making: a unifying computational neuroscience approach.* Oxford University Press, Oxford New York, 2008.

[265] Amir Rosenfeld, Mahdi Biparva, and John K. Tsotsos. Priming neural networks. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, June 2018.

[266] Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room, 2018.

[267] Ruth Rosenholtz. Visual search for orientation among heterogeneous distractors: Experimental results and implications for signal-detection theory models of search. *Journal of Experimental Psychology: Human Perception and Performance*, 27(4):985–999, 2001.

[268] Ruth Rosenholtz, Allen L. Nagy, and Nicole R. Bell. The effect of background color on asymmetries in color search. *Journal of Vision*, 4(3):9, mar 2004.

[269] Lars O. M. Rothkegel, Hans A. Trukenbrod, Heiko H. Schütt, Felix A. Wichmann, and Ralf Engbert. Temporal evolution of the central fixation bias in scene viewing. *Journal of Vision*, 17(13):3, nov 2017.

[270] Constantin A. Rothkopf, Dana H. Ballard, and Mary M. Hayhoe. Task and context determine where you look. *Journal of Vision*, 7(14):16, jul 2016.

[271] Guillaume A. Rousselet, Simon J. Thorpe, and Michèle Fabre-Thorpe. How parallel is visual processing in the ventral pathway? *Trends in Cognitive Sciences*, 8(8):363–370, August 2004.

[272] Edgar Rubin. *Figure and Ground.* Psychology Press, 1915/2001.

[273] David E. Rumelhart, Geoffrey E. Hinton, and Ronald J. Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, October 1986.

[274] Dov Sagi and Bela Julesz. Detection versus discrimination of visual orientation. *Perception*, 13(5):619–628, oct 1984.

[275] Dario D. Salvucci and Joseph H. Goldberg. Identifying fixations and saccades in eye-tracking protocols. In *Proceedings of the symposium on Eye tracking research & applications - ETRA '00*. ACM Press, 2000.

[276] Arthur G. Samuel and Donna Kat. Inhibition of return: A graphical meta-analysis of its time course and an empirical test of its temporal and spatial properties. *Psychonomic Bulletin & Review*, 10(4):897–906, dec 2003.

[277] M. G. Saslow. Latency for saccadic eye movement. *JOURNAL OF THE OPTICAL SOCIETY OF AMERICA*, 57(8), aug 1967.

[278] J.D. Schall. Frontal eye fields. In *Encyclopedia of Neuroscience*, pages 367–374. Elsevier, 2009.

[279] Boris Schauerte and Rainer Stiefelhagen. Quaternion-based spectral saliency detection for eye fixation prediction. In *Computer Vision – ECCV 2012*, pages 116–129. Springer Berlin Heidelberg, 2012.

[280] P. H. Schiller, J. H. Sandell, and J. H. Maunsell. The effect of frontal eye field and superior colliculus lesions on saccadic latencies in the rhesus monkey. *Journal of Neurophysiology*, 57(4):1033–1049, apr 1987.

[281] Peter H. Schiller and Edward J. Tehovnik. Chapter 9 look and see: how the brain moves your eyes about. In *Progress in Brain Research*, pages 127–142. Elsevier, 2001.

[282] Mark M. Schira, Christopher W. Tyler, Branka Spehar, and Michael Breakspear. Modeling magnification and anisotropy in the primate foveal confluence. *PLoS Computational Biology*, 6(1):e1000651, jan 2010.

[283] E. L. Schwartz. Spatial mapping in the primate sensory projection: Analytic structure and relevance to perception. *Biological Cybernetics*, 25(4):181–194, dec 1977.

[284] Al Seckel. *The ultimate book of optical illusions*. Sterling Pub. Co, New York, 2006.

[285] Nuno Seco, Tony Veale, and Jer Hayes. An intrinsic information content metric for semantic similarity in wordnet. In *ECAI*, 2004.

[286] H. J. Seo and P. Milanfar. Static and space-time visual saliency detection by self-resemblance. *Journal of Vision*, 9(12):15–15, nov 2009.

[287] Thomas Serre. Hierarchical models of the visual system. In *Encyclopedia of Computational Neuroscience*, pages 1–12. Springer New York, 2014.

[288] Thomas Serre, Gabriel Kreiman, Minjoon Kouh, Charles Cadieu, Ulf Knoblich, and Tomaso Poggio. A quantitative theory of immediate visual recognition. In *Progress in Brain Research*, pages 33–56. Elsevier, 2007.

[289] Thomas Serre, Lior Wolf, Stanley Bileschi, Maximilian Riesenhuber, and Tomaso Poggio. Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(3):411–426, March 2007.

[290] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, jul 1948.

[291] Robert Shapley and Michael J. Hawken. Color in the cortex: single- and double-opponent cells. *Vision Research*, 51(7):701–717, apr 2011.

[292] Bhavin R. Sheth and Ryan Young. Two visual pathways in primates based on sampling of space: Exploitation and exploration of visual information. *Frontiers in Integrative Neuroscience*, 10, nov 2016.

[293] Daniel J Simons and Christopher F Chabris. Gorillas in our midst: Sustained inattentional blindness for dynamic events. *Perception*, 28(9):1059–1074, September 1999.

[294] Lawrence C. Sincich and Jonathan C. Horton. THE CIRCUITRY OF v1 AND v2: Integration of color, form, and motion. *Annual Review of Neuroscience*, 28(1):303–326, jul 2005.

[295] SMI. *iViewX System Manual (IVX-2.4-0908)*. SensoMotoric Instruments GmbH, 2009.

[296] Samuel G. Solomon and Peter Lennie. The machinery of colour vision. *Nature Reviews Neuroscience*, 8(4):276–286, apr 2007.

[297] A. Soltani and C. Koch. Visual saliency computations: Mechanisms, constraints, and the effect of feedback. *Journal of Neuroscience*, 30(38):12831–12843, sep 2010.

[298] Marc A. Sommer. The spatial relationship between scanning saccades and express saccades. *Vision Research*, 37(19):2745–2756, oct 1997.

[299] Lothar Spillmann, Birgitta Dresp-Langley, and Chia huei Tseng. Beyond the classical receptive field: The effect of contextual stimuli. *Journal of Vision*, 15(9):7, July 2015.

[300] M.W. Spratling. Predictive coding as a model of the v1 saliency map hypothesis. *Neural Networks*, 26:7–28, feb 2012.

[301] Tania Stathaki. *Image fusion : algorithms and applications*. Academic Press/Elsevier, Amsterdam Boston, 2008.

[302] Andrew Stockman, Donald I. A. MacLeod, and Nancy E. Johnson. Spectral sensitivities of the human cones. *Journal of the Optical Society of America A*, 10(12):2491, dec 1993.

[303] H. Strasburger, I. Rentschler, and M. Juttner. Peripheral vision and pattern recognition: A review. *Journal of Vision*, 11(5):13–13, dec 2011.

[304] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. Sequence to sequence learning with neural networks, 2014.

[305] Benjamin W. Tatler. The central fixation bias in scene viewing: Selecting an optimal viewing position independently of motor biases and image feature distributions. *Journal of Vision*, 7(14):4, nov 2007.

[306] Benjamin W. Tatler, Roland J. Baddeley, and Iain D. Gilchrist. Visual correlates of fixation selection: effects of scale and time. *Vision Research*, 45(5):643–659, mar 2005.

[307] Benjamin W. Tatler, Roland J. Baddeley, and Benjamin T. Vincent. The long and the short of it: Spatial statistics at fixation vary with saccade amplitude and task. *Vision Research*, 46(12):1857–1862, jun 2006.

[308] Benjamin W Tatler, Nicholas J Wade, Hoi Kwan, John M Findlay, and Boris M Velichkovsky. Yarbus, eye movements, and vision. *i-Perception*, 1(1):7–27, January 2010.

[309] Amirhossein Tavanaei, Masoud Ghodrati, Saeed Reza Kheradpisheh, Timothée Masquelier, and Anthony Maida. Deep learning in spiking neural networks. *Neural Networks*, 111:47–63, March 2019.

[310] A. Tavassoli, I. van der Linde, A.C. Bovik, and L.K. Cormack. Eye movements selective for spatial frequency and orientation during active visual search. *Vision Research*, 49(2):173–181, jan 2009.

[311] Jan Theeuwes. Top-down search strategies cannot override attentional capture. *Psychonomic Bulletin & Review*, 11(1):65–70, feb 2004.

[312] A. Thielscher and H. Neumann. A computational model to link psychophysics and cortical cell activation patterns in human texture processing. *Journal of Computational Neuroscience*, 22(3):255–282, November 2006.

[313] Christopher Lee Thomas. Opensalicon: An open source implementation of the salicon saliency model. Technical Report TR-2016-02, University of Pittsburgh, 2016.

[314] Simon J. Thorpe and Michel Imbert. Biological constraints on connectionist modelling. 1989.

[315] Antonio Torralba and Aude Oliva. Statistics of natural image categories. *Network: Computation in Neural Systems*, 14(3):391–412, January 2003.

[316] Antonio Torralba, Aude Oliva, Monica S. Castelhano, and John M. Henderson. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological Review*, 113(4):766–786, 2006.

[317] Anne Treisman. Features and objects: The fourteenth bartlett memorial lecture. *The Quarterly Journal of Experimental Psychology Section A*, 40(2):201–237, may 1988.

[318] Anne Treisman. Search, similarity, and integration of features between and within dimensions. *Journal of Experimental Psychology: Human Perception and Performance*, 17(3):652–676, 1991.

[319] Anne Treisman and Stephen Gormican. Feature analysis in early vision: Evidence from search asymmetries. *Psychological Review*, 95(1):15–48, 1988.

[320] Anne Treisman and Janet Souther. Search asymmetry: A diagnostic for preattentive processing of separable features. *Journal of Experimental Psychology: General*, 114(3):285–310, 1985.

[321] Anne M. Treisman and Garry Gelade. A feature-integration theory of attention. *Cognitive Psychology*, 12(1):97–136, jan 1980.

[322] Colwyn B. Trevarthen. Two mechanisms of vision in primates. *Psychologische Forschung*, 31(4):299–337, 1968.

[323] Xoana Troncoso, Stephen Macknik, and Susana Martinez-Conde. Corner salience varies linearly with corner angle during flicker-augmented contrast: a general principle of corner perception based on vasarely's artworks. *Spatial Vision*, 22(3):211–224, may 2009.

[324] Xoana G Troncoso, Stephen L Macknik, and Susana Martinez-Conde. Novel visual illusions related to vasarely's 'nested squares' show that corner salience varies with corner angle. *Perception*, 34(4):409–420, apr 2005.

[325] John Tsotsos. *A computational perspective on visual attention*. MIT Press, Cambridge, Mass, 2011.

[326] John K. Tsotsos, Scan M. Culhane, Winky Yan Kei Wai, Yuzhong Lai, Neal Davis, and Fernando Nuflo. Modeling visual attention via selective tuning. *Artificial Intelligence*, 78(1-2):507–545, oct 1995.

[327] John K. Tsotsos, Iuliia Kotseruba, and Calden Wloka. A focus on selection for fixation. *Journal of Eye Movement Research*, 9(5):1–34, may 2016.

[328] John K. Tsotsos and Wouter Kruijne. Cognitive programs: software for attention's executive. *Frontiers in Psychology*, 5, nov 2014.

[329] E Tulving and D. Schacter. Priming and human memory systems. *Science*, 247(4940):301–306, January 1990.

[330] Eric Tzeng, Judy Hoffman, Trevor Darrell, and Kate Saenko. Simultaneous deep transfer across domains and tasks. In *2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, December 2015.

[331] Pieter J. A. Unema, Sebastian Pannasch, Markus Joos, and Boris M. Velichkovsky. Time course of information processing during scene perception: The relationship between saccade amplitude and fixation duration. *Visual Cognition*, 12(3):473–494, apr 2005.

[332] Wieske van Zoest and Mieke Donk. Saccadic target selection as a function of time. *Spatial Vision*, 19(1):61–76, jan 2006.

[333] David C. VanEssen and Jack L. Gallant. Neural mechanisms of form and motion processing in the primate visual system. *Neuron*, 13(1):1–10, jul 1994.

[334] Rufin VanRullen. Visual saliency and spike timing in the ventral visual pathway. *Journal of Physiology-Paris*, 97(2-3):365–377, March 2003.

[335] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need, 2017.

[336] Richard Veale, Ziad M. Hafed, and Masatoshi Yoshida. How is visual salience computed in the brain? insights from behaviour, neurobiology and modelling. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 372(1714):20160113, jan 2017.

[337] Preeti Verghese. Visual search and attention: A signal detection theory approach. *Neuron*, 31(4):523–535, aug 2001.

[338] Benjamin T. Vincent and Benjamin W. Tatler. Systematic tendencies in scene viewing, 2008.

[339] Melissa L.-H. Võ and John M. Henderson. Object–scene inconsistencies do not capture gaze: evidence from the flash-preview moving-window paradigm. *Attention, Perception, & Psychophysics*, 73(6):1742–1753, may 2011.

[340] Nicholas Wade. *The art and science of visual illusions*. Routledge & Kegan Paul, London Boston, 1982.

[341] Brian Wandell. *Foundations of vision*. Sinauer Associates, Sunderland, Mass, 1995.

[342] Brian A. Wandell, Serge O. Dumoulin, and Alyssa A. Brewer. Visual field maps in human cortex. *Neuron*, 56(2):366–383, oct 2007.

[343] C.-A. Wang, S. E. Boehnke, L. Itti, and D. P. Munoz. Transient pupil response is modulated by contrast-based saliency. *Journal of Neuroscience*, 34(2):408–417, jan 2014.

[344] Yixiu Wang, Bin Wang, Xiaofeng Wu, and Liming Zhang. Scanpath estimation based on foveated image saliency. *Cognitive Processing*, 18(1):87–95, oct 2016.

[345] A. B. Watson. A formula for human retinal ganglion cell receptive field density as a function of visual field location. *Journal of Vision*, 14(7):15–15, jun 2014.

[346] Hui Wei and Zheng Dong. Contour representation and shape matching based on mechanism of visual cortex. In *2016 International Joint Conference on Neural Networks (IJCNN)*. IEEE, July 2016.

[347] Michael Weliky, Karl Kandler, David Fitzpatrick, and Lawrence C. Katz. Patterns of excitation and inhibition evoked by horizontal connections in visual cortex share a common relationship to orientation columns. *Neuron*, 15(3):541–552, sep 1995.

[348] John Werner and Leo M. Chalupa. *The new visual neurosciences*. The MIT Press, Cambridge, Massachusetts, 2014.

[349] M. Wertheimer. *Laws of organization in perceptual forms*. Harcourt, Brace & Jovanovitch, London, 1923/1938.

[350] Brian White and Douglas P. Munoz. *The Oxford Handbook of Eye Movements*. Oxford University Press, aug 2011.

[351] Brian J. White, David J. Berg, Janis Y. Kan, Robert A. Marino, Laurent Itti, and Douglas P. Munoz. Superior colliculus neurons encode a visual saliency map during free viewing of natural dynamic video. *Nature Communications*, 8:14263, jan 2017.

[352] Brian J. White, Janis Y. Kan, Ron Levy, Laurent Itti, and Douglas P. Munoz. Superior colliculus encodes visual saliency before the primary visual cortex. *Proceedings of the National Academy of Sciences*, 114(35):9451–9456, aug 2017.

[353] Michael White. A new effect of pattern on perceived lightness. *Perception*, 8(4):413–416, aug 1979.

[354] Gagan S Wig, Scott T Grafton, Kathryn E Demos, and William M Kelley. Reductions in neural activity underlie behavioral components of repetition priming. *Nature Neuroscience*, 8(9):1228–1233, July 2005.

[355] Stefan Winkler and Ramanathan Subramanian. Overview of eye tracking datasets. In *2013 Fifth International Workshop on Quality of Multimedia Experience (QoMEX)*. IEEE, jul 2013.

[356] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Saccade sequence prediction: Beyond static saliency maps, 2017.

[357] Calden Wloka, Iuliia Kotseruba, and John K. Tsotsos. Active fixation control to predict saccade sequences. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*. IEEE, jun 2018.

[358] Calden Wloka and John Tsotsos. Spatially binned ROC: A comprehensive saliency metric. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, jun 2016.

[359] J. M. Wolfe. Guided search 4.0: A guided search model that does not require memory for rejected distractors. *Journal of Vision*, 1(3):349–349, mar 2010.

[360] Jeremy Wolfe. *Attention*. Psychology Press, 1998.

[361] Jeremy M. Wolfe. “effortless” texture segmentation and “parallel” visual search are not the same thing. *Vision Research*, 32(4):757–763, apr 1992.

[362] Jeremy M. Wolfe. Guided search 2.0 a revised model of visual search. *Psychonomic Bulletin & Review*, 1(2):202–238, jun 1994.

[363] Jeremy M. Wolfe. Asymmetries in visual search: An introduction. *Perception & Psychophysics*, 63(3):381–389, apr 2001.

[364] Jeremy M. Wolfe and Sara C. Bennett. Preattentive object files: Shapeless bundles of basic features. *Vision Research*, 37(1):25–43, jan 1997.

[365] Jeremy M. Wolfe, Kyle R. Cave, and Susan L. Franzel. Guided search: An alternative to the feature integration model for visual search. *Journal of Experimental Psychology: Human Perception and Performance*, 15(3):419–433, 1989.

[366] Jeremy M. Wolfe and Todd S. Horowitz. What attributes guide the deployment of visual attention and how do they do it? *Nature Reviews Neuroscience*, 5(6):495–501, jun 2004.

[367] Jeremy M. Wolfe, Evan M. Palmer, and Todd S. Horowitz. Reaction time distributions constrain models of visual search. *Vision Research*, 50(14):1304–1311, jun 2010.

[368] Jeremy M. Wolfe, Ester Reijnen, Todd S. Horowitz, Riccardo Pedersini, Yair Pinto, and Johan Hulleman. How does our search engine “see” the world? the case of amodal completion. *Attention, Perception, & Psychophysics*, 73(4):1054–1064, feb 2011.

[369] S.Sabina Wolfson and Michael S. Landy. Discrimination of orientation-defined texture edges. *Vision Research*, 35(20):2863–2877, oct 1995.

[370] Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *Proceedings of the 32nd annual meeting on Association for Computational Linguistics -*. Association for Computational Linguistics, 1994.

[371] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention, 2015.

[372] Yin Yan, Li Zhaoping, and Wu Li. Bottom-up saliency and top-down learning in the primary visual cortex of monkeys. *Proceedings of the National Academy of Sciences*, page 201803854, sep 2018.

[373] Alfred L. Yarbus. *Eye Movements and Vision*. Springer US, 1967.

[374] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 27*, pages 3320–3328. Curran Associates, Inc., 2014.

[375] Laure Zago, Mark J. Fenske, Elissa Aminoff, and Moshe Bar. The rise and fall of priming: How visual exposure shapes cortical representations of objects. *Cerebral Cortex*, 15(11):1655–1665, February 2005.

[376] Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In *Computer Vision – ECCV 2014*, pages 818–833. Springer International Publishing, 2014.

[377] Jianming Zhang and Stan Sclaroff. Saliency detection: A boolean map approach. In *2013 IEEE International Conference on Computer Vision*. IEEE, dec 2013.

[378] Jianming Zhang and Stan Sclaroff. Exploiting surroundedness for saliency detection: A boolean map approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(5):889–902, may 2016.

[379] Liming Zhang and Weisi Lin. *Selective Visual Attention*. John Wiley & Sons (Asia) Pte Ltd, mar 2013.

[380] Lingyun Zhang, Matthew H. Tong, Tim K. Marks, Honghao Shan, and Garrison W. Cottrell. SUN: A bayesian framework for saliency using natural statistics. *Journal of Vision*, 8(7):32, dec 2008.

[381] L. Zhaoping. Gaze capture by eye-of-origin singletons: Interdependence with awareness. *Journal of Vision*, 12(2):17–17, feb 2012.

[382] Li Zhaoping. V1 mechanisms and some figure–ground and border effects. *Journal of Physiology-Paris*, 97(4-6):503–515, jul 2003.

[383] Li Zhaoping. *Understanding vision : theory, models, and data*. Oxford University Press, Oxford, United Kingdom, 2014.

[384] Li Zhaoping. From the optic tectum to the primary visual cortex: migration through evolution of the saliency map for exogenous attentional guidance. *Current Opinion in Neurobiology*, 40:94–102, oct 2016.

[385] Li Zhaoping and Keith A. May. Psychophysical tests of the hypothesis of a bottom-up saliency map in primary visual cortex. *PLoS Computational Biology*, 3(4):e62, 2007.

[386] Li Zhaoping and Li Zhe. Primary visual cortex as a saliency map: A parameter-free prediction and its test by behavioral data. *PLOS Computational Biology*, 11(10):e1004375, oct 2015.

[387] Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. Learning deep features for discriminative localization. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, June 2016.