MANCHESTER 1824
The University of Manchester

A bio-inspired attentive model to predict perceptual saliency in natural scenes

A thesis submitted to the University of Manchester for the degree of
Doctor of Philosophy
in the Faculty of Science and Engineering

2022

Giulia D'Angelo

Supervisors:

Dr. Chiara Bartolozzi - Istituto Italiano di Tecnologia

Dr. Vadim Tikhanoff - Istituto Italiano di Tecnologia

Prof. Angelo Cangelosi - The University of Manchester

Department of Computer Science, School of Engineering

Contents

List of figures

List of publications

- Iacono, M., D'Angelo, G., Glover, A., Tikhanoff, V., Niebur, E., & Bartolozzi, C. (2019, November). Proto-object based saliency for event-driven cameras. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 805-812). IEEE.
- D'Angelo, G., Janotte, E., Schoepe, T., O'Keeffe, J., Milde, M. B., Chicca, E., & Bartolozzi, C. (2020). Event-based eccentric motion detection exploiting time difference encoding. Frontiers in neuroscience, 14, 451.
- Ghosh, S., D'Angelo, G., Glover, A., Iacono, M., Niebur, E., & Bartolozzi, C. (2022). Event-driven proto-object based saliency in 3D space to attract a robot's attention. Scientific reports, 12(1), 1-14.
- D'Angelo, G., Perrett, A., Iacono, M., Furber, S., & Bartolozzi, C. (2022). Event driven bio-inspired attentive system for the iCub humanoid robot on SpiNNaker. Neuromorphic Computing and Engineering, 2(2), 024008.

Terms and abbreviations

AER	Address-Event Representation
ATIS	Asynchronous Time-based Image Sensor
AUC	Area under ROC Curve
BO	Border Ownership
CC	Pearson's Correlation Coefficient
CPU	Central Processing Unit
CS	Center Surround
CSP	Center Surround Pyramids
DART	Distribution Aware Retinal Transform
DVS	Dynamic Vision Sensor
DYNAPs	Dynamic Neuromorphic Asynchronous Processors
evProto	event-driven Proto-object model (Chapter 2)
evProtoDepth	event-driven Proto-object model Depth (Chapter 4)
fbProtoDepth	frame-based Proto-object model Depth (Chapter 4)
GPU	Graphics Processing Unit
MSE	Mean squared error
NSS	Normalized Scanpath Saliency
OL	Overlapping Percentaage
RF	Receptive Field
ROI	Regions of Interest
ROC	Receiver operating characteristic
sEMD	spiking Elementary Motion Detector
SNN	Spiking Neural Network
SSIM	Structural Similarity
VM	Von Mises
V2	Secondary Visual Cortex
WTA	Winner-Take-All

Abstract

The sensory input from the entire visual field carried through the optic nerve to the visual system could exceed the processing capabilities of the cortex. Focalisation towards specific areas of interest represents the natural coping mechanism of processing exclusively the relevant part of the visual field. State-of-the-art mainstream artificial vision, relying on frame-based cameras and convolutional neural networks, exploits the current availability of computational resources but falls short when the visual system has to be deployed in compact, autonomous systems that cannot rely on external access to computing devices. This leads to the following questions: Can we reduce computational load via bioinspired visual attention mechanisms? Can a robot take advantage of the same attentional mechanisms to quickly interact with the environment? This work addresses these questions by bridging the gap between biologically inspired vision sensors and models of attention and resulted in the implementation of an event-driven model of attention on the humanoid robot iCub.

Biologically inspired event-driven vision sensors are loosely inspired by the retina parvo magno-parvo and magnocellular pathway, reacting to changes in the field of view. They reduce the redundancy of the visual signal related to static stimuli and produce a stream of spikes that encode information similarly to biological neurons. Biologically inspired models of visual attention explain which mechanisms drive the selection of salient stimuli in the visual input. To test the assumption that biologically inspired vision sensors coupled with attention models can be exploited to select relevant stimuli for a robot, I selected three main event-driven bottom-up feature extraction channels fed into a biologically plausible saliency model based on the Gestalt theory of perceptual grouping. Intensity, disparity and motion are the first information cues of this project towards a more complex attention model where the channels compete with each other to represent the scene.

My work demonstrated the applicability of biologically plausible event-driven saliency-based visual attention models for iCub. These models can run online and on neuromorphic platforms proving the possibility of exploiting fully bioinspired pipelines to determine visual attention cues with low latency.

Declaration of originality

I hereby confirm that no portion of the work referred to in the thesis has been submitted in support of an application for another degree or qualification of this or any other university or other institute of learning.

Copyright statement

Acknowledgements

How can I sum up these three years?
Maybe there is no need for a nice wrap-up of what happened, maybe I can just bring with me little moments of rare gratitude and happiness as long as challenging periods where I have grown the most. The constant feeling of not being enough took me to the idea that I will never fit perfectly in, and that I am perfectly imperfect as I am. I will never know everything necessary to impress others, but I will always keep trying to impress myself with my inexhaustible passion.
I always thought that people make a difference, and they made it. I am extremely lucky, and these words cannot give justice to the gratitude I feel.
Chiara has been everything I could have ever asked for. She is my mentor, a woman I take inspiration from, a brilliant scientist and a friend. She has always been by my side, encouraging me to push the limits I forced myself into because of fear. I will never thank her enough for the personal and professional growth she helped me to gain.
I met Vadim right after the university. I will never forget he has been the first one to believe in me. Working with him never feels like real work. He is caring and extremely good at lifting you. He has always treated me as a colleague leaving me room to speak without judgement.
I still remember Angelo's words during an annual report, "Giulia, why did you stop smiling after the question? it's ok, respond with no fear, everything it's fine". Angelo is an empathetic person who never missed a chance to be understanding.
The person I will never stop thanking for being right next to me every day is Massi. My best friend, my PostDoc guide, my personal smile and the perfect hug when doubts take over. The gratitude I have for having him in my life is indescribable with words.
My family, my mother and my siblings are the owners of my heart. Their love touches me and builds my strength. They were, are and will be my world.
My life is full of wonderful friends who represent the family I chose.
Beck is my unstoppable source of laughs and understanding, Elisa is the strong and caring woman I wish to become, Ella is my buddy, a smile everywhere we are, Luna is all the strength I need when I'm overwhelmed, Martino a boost of serenity, Suman the inexorable worker and ally, Marco and Fabrizio my daily support, Adam a discovered friend and Enea a truly pure friend.

Nico, Gianma, Damiano and Ruzze are my nerdy safe space.
Sara, Giorgia and Silvia my bundle of joy, Stefano my certainty, Elisa my dose of cuddles, Mattia and Andrea the rhythm of my laughs.
Silvia and Federica know how to make dancing my heart and soul.
A particular thought goes to Jay, who has been right next to me since we met, my personal angel.
Without all of these people, I could not have had the happiness my life reserved for me. Thanks, from every single part of me.
And thanks to me. Thanks for allowing me to be surprised by my spontaneous curiosity. Thanks for having always imagined a new light after every storm.
Ciao papà, c'è un po' di te in tutto questo, ti penso.

"The brain is imagination, and that was exciting to me; I wanted to build a chip that could imagine something"

Misha Mahowald

Chapter 1

Introduction

1.1 Motivation

This work seeks inspiration from the complex mechanisms nature has found to solve daily tasks throughout evolution. To make decisions for its life, an agent needs to process the information about the surroundings through perception. A sudden smell, sound or movement could lead an agent to escape from predators or look for food. In particular, the majority of mammals rely on the visual system to sense the environment [3]. Although the visual system is a complex structure where sophisticated mechanisms work to perceive the scene, processing all the external stimuli is too computationally demanding. Attention allows focusing on a specific area elaborating only on what is ”important” for a specific task [4].
A robot working in an unconstrained scenario can take advantage of attention mechanisms exploiting bioinspired algorithms to reduce computational loads and latency. Specifically, this work wants to exploit smart solutions such as visual attention mechanisms allowing a ”natural” and fast response of the robot to external stimuli. The aim of my work is to create a fully bio-inspired pipeline for the humanoid robot iCub [2](see Figure 1.1) by bridging biologically plausible models of attention with biologically inspired (or ”neuromorphic”) hardware for sensing and processing. The synergy between neuromorphic algorithms and bioinspired hardware is the focal contribution of this project, made possible thanks to a pair of a new generation of neuromorphic cameras [5] mounted on the iCub platform and processing hardware [6] providing the perfect basis to explore spike-based models. In the brain, neurons communicate via neuron spikes (action potentials) transmitting the information through spike trains thanks to the synaptic connections. This thesis is a starting point for more complex bio-inspired attention models seeking to prove the reliability of fully biologically plausible systems reducing the amount of data to be processed and the latency to obtain a response from the model, supporting the development of artificial systems capable of swift interaction within an unconstrained environment.

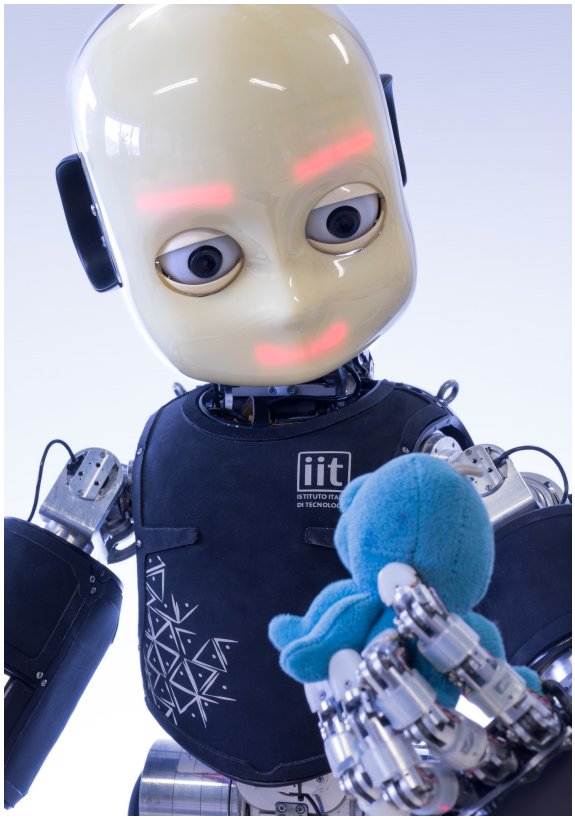


Figure 1.1. Figure of iCub, the humanoid robot focusing on an object. iCub is one of the main research platforms used in embodied cognition research [1], it has been created for research purposes, it is 104 cm tall with sophisticated hands and in the form of a child. One version of the robot is the neuromorphic iCub [2], equipped with bio-inspired cameras and event-driven skin. (Further information https://icub.iit.it/)

1.2 Scope

Nowadays, the possibility to live in a world where robots are among us starts to be a reality. Robots can autonomously solve daily tiring tasks improving our lives. The intent behind this work is to exploit visual attention mechanisms to allow a robot to perceive external stimuli focusing only on salient regions of the scene, granting a response in a time comparable to perform a saccade (200ms [7], [8]). To do so, I explored the opportunity to tailor models of attention based on computational principles observed in the brain, with neuromorphic sensing and processing hardware. Despite being inspired by biology, the attention model inspiring this project relies on an artificial input, i.e. static images that are very far from the signals the brain receives. The motion detector instead encodes information in a way that is closer to biology, relying on spikes and their spatiotemporal distribution. The real problem this project faced is understanding if biologically plausible saliency-based attention models could realistically take part in a complex attention schema for a humanoid robot.

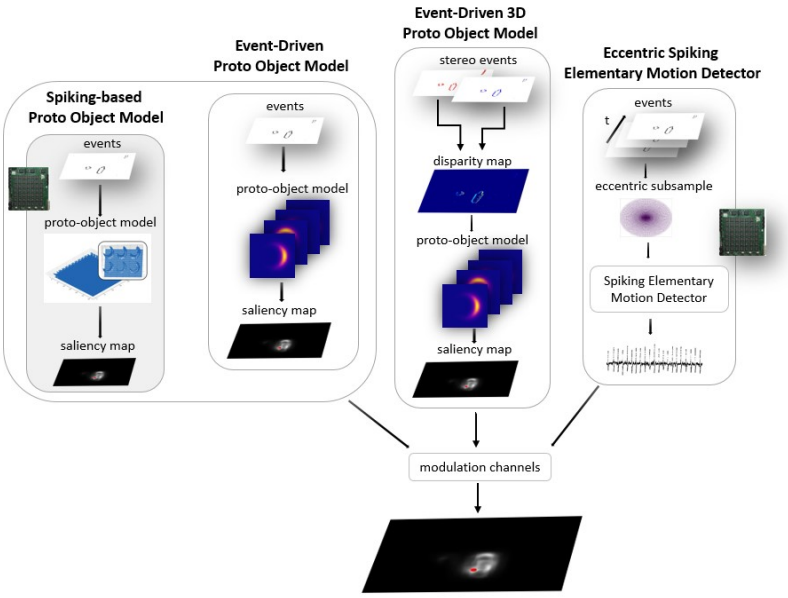


Figure 1.2. Schematic representation of the three channels: intensity (Event-Driven Proto Object model and Spiking-based Proto Object model), depth (Event-Driven 3D Proto Object model) and motion (Eccentric Spiking Elementary Motion Detector). The contribution of each channel will be modulated to obtain the final saliency map.

iCub is an open-source research platform designed to test the embodied cognition hypothesis where the internal model of the world is strongly determined by the form of the body. Developing an attention-based model on iCub allows easy integration with bioinspired algorithms and hardware. It may be expected that, the use of bioinspired and neuromorphic hardware such as SpiNNaker [6] and the ATIS sensors [5] in combination with biological pipelines [9]–[11] should lead to a reduction of data load with a consequential reduction in power consumption and computational time allowing a quick and ”natural” response from the robot [12].

Neuromorphic hardware does not guarantee the scalability of big complex networks with a high number of neurons demanding for a challenging scientific question. In recent years bioinspired and neuromorphic systems with extreme low latency have proven to be suitable for several different applications [13]–[18]. These bioinspired systems have not yet been explored for a biologically inspired pipeline building an attention model for a humanoid robot like iCub.

1.3 Thesis Outline

The world is full of visual stimuli needing for a quick response from agents to allow interaction or avoidance. The urgency for robots, living in an unconstrained environment,

to be attentive to their surrounding becomes crucial. Visual attention mechanisms are the core of this project to the basis of the human selection of salient regions (Ch. 1.4). I have focused my work on saliency-driven (bottom-up) models of attention. These models reproduce non-bioinspired and bioinspired models to obtain a saliency map of the scene (Ch. 1.4.1). The originally proposed models were purely feature-based disregarding the feature integration as a key role in attention [19] further missing human perceptual grouping theories. Human beings instinctively group items in the scene following perceptual organisation theories introduced by the Gestalt principles (Ch. 1.4.2). These theories clearly illustrate the importance of border ownership in perception subsequently demonstrated as existent mechanisms in the Secondary Visual Cortex (see Ch. 1.4.3). The Gestalt theories and the presence of Border Ownership cells in the cortex inspired scientists towards biologically plausible saliency-based proto-object attention models, defining a *proto-object* as an area of the visual field where potentially there is an object (Ch. 1.4.4).
This Thesis project relies on the implementation of a fully bioinspired attention system on neuromorphic hardware (Ch. 1.5) where different channels of information pursue the research of salient regions of the scene. These three channels of information are intensity (see Ch. 2 and Ch. 3), depth (see Ch. 4) and motion (see Ch. 5). Each chapter describes a single channel showing results, performances and limitations of the resulting attention model. These channels represent the first attempt towards a more complex attention model where the saliency map is given by the interpretation and modulation of these parallel channels of information cues (see Figure 1.2).

1.4 Attention Mechanisms

'Everyone knows what attention is. It is the taking possession by the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration, of consciousness are of its essence. It implies withdrawal from some things in order to deal effectively with others' [20]. Agents, be they biological (animals) or technological (robots) thrown into a complex environment need to organise their sensory input in an efficient way, to allow efficient exploration of their surroundings and exploitation of their resources. The exploration of the environment to interact with the surroundings implicitly involves the mechanism of visual-spatial attention and alertness.
Visual attention can be subdivided into intensive and selective phenomena. The state of alertness or arousal can be defined as an intensive phenomenon, while the ability to orient visual attention to a particular region of the scene is described as selective mechanism. Since the agent's computational capabilities are limited, this requires careful allocation of

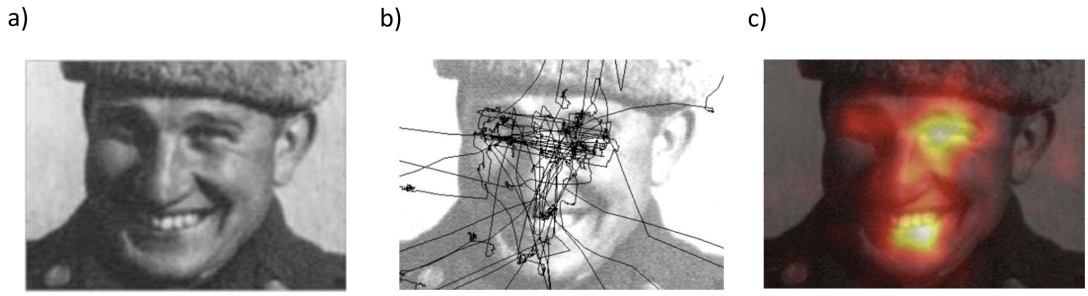


Figure 1.3. Scan path and fixation map. a) Photo of the 20th century psychologist Alfred Yarbus shown to human observers. b) Trace of eye movements (scan path) from one observer overlaid onto the image. Note frequent dwelling of gaze on eyes and mouth. c) Fixation map of the scene. The intensity of the coloured overlay represents the amount of fixation time received across all observers. All three figures are adapted from Tatler et. al.[23] licensed under CC BY-NC-ND 3.0.

perceptual and cognitive resources [21]. The study of this process goes back to antiquity when, in the 5th century AD, Augustine of Hippo studied how attention is attracted by different events [22].
The selection of interesting regions of the scene typically leads to the shift of attention towards the target.

The mechanism covert of mental attention shift without a physical eye movement is known as covert attention, whilst directing the eyes towards the salient point is called overt attention. The selection of interesting items is still a complex open topic. Overt attention has been explored focusing the research on the mechanisms behind the human saccades towards a particular target.
Substantial progress was made in the 1960s by Alfred Yarbus [24] who constructed a rudimentary eye tracker that made it possible to record the eye movements of subjects looking at paintings, photographs, etc. The aim, yet again, was to discern what attracts attention and therefore, to determine which parts of a scene humans fixate on. The outcome emphasised the complexity of this topic: attention is a result of the complex interplay between the extraction of features from the scene as well as task-driven mechanisms. In some cases, the subjects were attracted by the edges of a profile, in others by the animal in the distance, and in others by the eyes and the mouth of a person. Overall, the subjects were interested in regions of the picture that were explanatory of the scene, giving more information to the observer [25]. For the stimuli that Yarbus chose, the gaze patterns depended heavily on the task (”top-down attention”) and, to a lesser extent, on parts of the scene that were characterised by their intrinsic salience, defined as being low-level stimulus characteristics without high-level semantic meaning (”bottom-up attention”). In all cases, images were not processed as structure-less collections of visual features; instead, visual scenes are organised in terms of objects, and attention can be directed preferentially to them [26].

The trajectory of eye movements is called the scan path for a given picture and for one observer. Averaging them either over multiple scan paths of one observer or over those of multiple observers, or both defines the mean scan path that is used to draw conclusions about attentional processes. Points of the scene where the observer was focused on a particular target determine a fixation. The collection of all the fixations is then called a fixation map, (see Figure 1.3). The amplitude in each point of this map is proportional to the time spent during fixation at the corresponding point of the image, i.e. the intensity of overt attention to this point. It has been known for more than a century that overt attention can be distinguished from covert attention, defined as moving attention directed towards different parts of the visual input without moving the centre of gaze towards them [27]. Although overt and covert attention can be dissociated [28], [29], they are generally assumed to be strongly and positively correlated during normal viewing [30]–[33]. Parkhurst et. al.[34] therefore suggested evaluating computational models of covert attention by testing their predictions against observed eye movements, a convenient procedure which has become the standard method in the field [35]. The shift of the gaze depends on internal desires/personal goals or external stimuli [36]. The bottom-up process extracts features from the scene perceptually organising alerting the agent of possible salient items. Top-down mechanisms come to play in modulating the bottom-up signals when the task is clearly defined [37]. A robot can similarly take advantage of basic attention behaviour as a paradigm to interact with the surroundings. Furthermore, it can be the foundation for more complex real-time behaviours, combining alertness with actions in response to an incoming stimulus, or simply reaching interesting items in the surrounding.

The evaluation of saliency-based models via human fixation maps is not only important behavioural research to emulate intelligent human-like eye movements that can improve the naturalness in social-robot interaction [38][39], but it is also important to orient the robot towards regions of the visual scene that are likely to be relevant in downstream tasks like segmentation, tracking and subsequent interaction.

Attention in robotics scenarios has been already tackled, providing perception to robots for social interaction tasks with humans [40] directing the robotic gaze towards specific regions of the scene using a face detector, colour and motion cues. This model attempts to combine top-down and bottom-up cues generating a map influenced by the habituation function allowing the robot to switch from one task to another one when it is *habituated* to a specific task. This model proposes a set of possible chosen ”Strategies” defining a priori a set of possible behaviours precluding an autonomous unconstrained exploration and interaction with the surrounding.

Attention mechanisms are exploited also to distribute processing from different feature extractions to provide a real-time response, moving a humanoid robot head, linearly com-

bining several conspicuity maps to generate a final saliency map [41]. This bottom-up model also claims top-down effects weighting the contribution of the conspicuity maps. The focus of this model is the parallel processing using a computer cluster architecture with 8 PCs that allows for proper distribution of the high computational load of the model. Similar attention approaches have been applied using saliency maps for an anthropomorphic robot [42]. This work, focused on the interplay between the oculomotor control and the visual processing integrating cues from vision, audition, haptics, and also top-down volitional inputs to allow overt visual attention. The sensory processing module feeds into the motor planning block subsequently generating a motor command. The system includes an 'Interaction Issue' computation maintaining the frame of reference once the robot's eyes move towards a target. In this work, spatially localised stimuli compete to become the next saccade target using a Winner-Take-All (WTA) mechanism representing the salient target worthy of attention.

A different work proposes a bottom-up attention system moving iCub's eyes based on visual and acoustic saliency maps exploiting an inhibition-of-return mechanism allowing the robot to explore seeking new salient regions [43].

The model computes saliency pre-filtering the image input extracting fundamental visual features (intensity, colour, motion and hue) and detecting the location of the sound source. The saliency map is then generated extracting the maximum value across all saliency channels at each location. This model moves iCub's eyes only using exploratory behaviour without taking into account a target task or situation-driven behaviour.

Another bioinspired model proposed for iCub combines orientation, contrast and flicker are combined over different scales, producing a saliency map allowing the robot to focus its attention on salient regions of the scene [44]. This model exploits the low latency due to bio-inspired cameras mounted on iCub dramatically decreasing the amount of data to be processed thanks to the reduction of redundant information. Furthermore, the system takes advantage of the attention system to detect interesting parts of the scene processing only relevant regions rather than the whole field of view. A saliency-based method significantly decreases the number of further computations, processing only interesting parts of the scene instead of analysing the entire visual field.

None of the mentioned attention models takes into account depth perception as an important feature to determine interaction with the environment. Depth plays a key role in the interpretation of the scene, and vice versa attention takes part in the initiation of the three-dimensional interpretation [45]. Furthermore, depth perception allows robots to safely explore the environment. These models are characterised by high computational loads and lack of important aspects for robotic attentional systems such as depth perception to reach specific targets or running online on the robot for real-time responses. Already in

1988 Clark et al. proposed a saliency-based motion control system to fixate specific 3D locations using depth as a further feature of the model [46]. Years later, Pahlavan et al. focused on the 'fixation vergence control' problem stating which portion of the image a robot should fixate on [47]. Moreover, in 2005, Neil and Tsotsos incorporated binocular disparity in a selective turning attention system prioritising features using biases [48]. Another work from Pasquale et al. has studied attention exploiting disparity-based segmentation using frame-based cameras on the iCub humanoid robot [49]. Motion is another important cue for attention mechanisms, either in selective attention tracking an object or avoiding an obstacle in an alert state. Motion is strongly modulated by attention mechanisms, especially in actions perception and recognition [50]. Motion perception has proven to shift the focus of attention during attentive tracking [51] and attract attention even when the motion per se is not informative [52]. Motion attention models have been proposed for video skimming [53], video captioning [54] or to predict human motion exploiting a feed-forward network comparing motion sub-sequences. Li et al. proposed an interesting pipeline for visual salient object detection using motion as a cue [55]. Motion attention algorithms have proven to be effective also for robotic applications where robotic movement depends on subjects' attention [56]. What seems missing in the literature is a bottom-up saliency-based attention schema exploiting a fully bioinspired pipeline where different cues cooperate to provide saliency in natural scenes seeking low latency allowing a "natural" response from humanoid robots.

1.4.1 Saliency-based models

During decades of attention studies, several attention models have been proposed and exploited for their characteristics to reduce the computational load focusing the processing only on salient regions of the scene. Saliency-based models share three main processes: the extraction of the features, the computation of the individual feature maps and the integration and/or competition to generate the final saliency map of the scene [57]. The saliency map can be used for various purposes, such as visual searching for selective attention [58], object recognition [59], tracking objects [60] proving the applicability of saliency-based models to define Regions of Interests (ROIs) for further localised processing. The work described in [60] in particular, is based on a recurrent neural network outperforming classical convolutional networks without the need to train the network.
Na Tong et al. [61] instead, proposed a saliency-based model tackling the noise reduction problem by generating two different saliency maps, weak and strong. This model learns the saliency via bootstrap learning including the centre bias phenomenon [62]. Attention models have also been proposed for human fixation prediction exploiting deep neural net-

works to predict saliency [63]. The saliency-based approach has been exploited also for robotic applications [64], [65] where iCub moves eyes and neck depending on bottom-up saliency multimodal model with visual and acoustic cues [43]. Given the complexity of these models, a fast approximation for visual saliency by Butko et al. [66] was proposed to reduce the heavy computations in saliency-based approaches to quickly orient robotic cameras toward human faces. A reinforcment learning saliency-based algorithm was proposed to guide an unmanned aerial vehicle through obstacle avoidance [67]. During years of attention studies, bioinspired saliency-based models have been developed exploiting bioinspired mechanisms to emulate neurons for visual attention tasks [68]–[72]. Mechanisms of selective attention have been further implemented by exploiting the benefit of neuromorphic circuits [73], [74]. One of the cornerstones of these models has been proposed by Itti and Koch [33]. This model extracts features (colour, intensity and orientation) from the visual input. These three channels of information compete with each other exploiting the WTA mechanism to represent the scene producing a final saliency map. Later on, the same authors describe five trends arisen in computational models of visual attention: saliency critically depends on the context, a unique 'saliency map' is an efficient and plausible bottom-up control strategy, inhibiting currently attended location (inhibition of return) is crucial for attentional deployment, attention and eye movements work together and the understanding of the scene and object recognition seem to constrain the selection of attended locations [75]. These models extract data from the scene disregarding the integration within features ruling the perceptual organisation of items in the visual field.

1.4.2 Gestalt laws

"The whole is other than the sum of the parts." This sentence, by Kurt Koffka, represents the meaning behind the Gestalt theory born in the early 20th century [76]. Gestalt principles try to explain how human beings perceive the world reducing complex scenes into simple shapes, counteracting the dominant structuralist view. The structuralism, described by the work of Wilhelm Wundt, Edward B. Titchener and Hermann von Helmholtz [77],[78], conceives the idea that complex items, ideas or thoughts, therefore any complex matter is always built from simple elements. Gestalt is a German word for "form" and "shape" and is used nowadays to describe the way a thing "has been put together". There is no exact equivalent in English. In psychology, the word is often interpreted as "pattern" or "configuration". Max Wertheimer, Kurt Koffka and Wolfgang Kohler refer to the word "Gestalt" also as "unified whole" devising the Gestalt principles for the first time.
The eyes perceive shapes, a multitude of elements as a whole, perceptually grouping items in the scene. The intrinsic content of an item, whether it is perceived or imagined, is de-

scribed from a set of sensations and images associated with that specific object. Therefore, the perception of items is the result of the sensations together with the images representing the item. The perceptions of shapes and patterns begin with fundamental local sensations. These shapes and local sensations work together to create the mental representation of the object. Helmholtz suggested that the idea we have in mind of an item carries on the information about the structure. For example, the structural idea we have of a table is built exploiting what we would expect to see from another viewpoint [79]. The *Gestalt theory*, suggests the perception as an effect of the stimulus configuration, disregarding previous hypotheses suggesting the perception of an item as an aggregation of local stimulus properties. In opposition to this theory, Hearing [80] and Mach [81] previously debated the perception of an item as due to specific neural interactions at the lowest level of the sensory system directly perceiving the properties. Ehrenfels [82], the author of the *Gestalt qualities*, finally explained the attribute of the whole configuration, adding the *Gestalt qualities* in the list of sensory primitives. The perception of an object is not only the combination of simple elements, its visual form is perceived also changing size, colour, orientation and etc.
Gestalt theory involves a series of different laws describing our ability to perceive the surroundings finding order in the disorder of the external stimuli. Four of these principles are known to be the most famous ones:

- **Closure (Reification) or Continuity**: We automatically fill gaps when looking at elements to perceive the complete image searching for the whole.
- **Common Region**: Elements belonging to the same closed region are grouped.
- **Figure-Ground**: We search for stable items to identify a figure segmenting the foreground from the background.
- **Proximity**: We group items based on their distance from each other. Items next to each other are more likely to be grouped together.

These mechanisms contribute together to the grouping of visual features into coherent objects [83]. Gestalt psychology further introduced the concept of border ownership in perceptual organisation highlighting the importance of discriminating items from the background.
These theories inspired scientists exploiting Gestalt laws to build models for robotic vision [84]–[86]. In the last decade, more recent work has been focused on the border ownership perceptual organisation using kernels to model feature extraction [9], [87]–[89] and figure-ground segmentation [90].

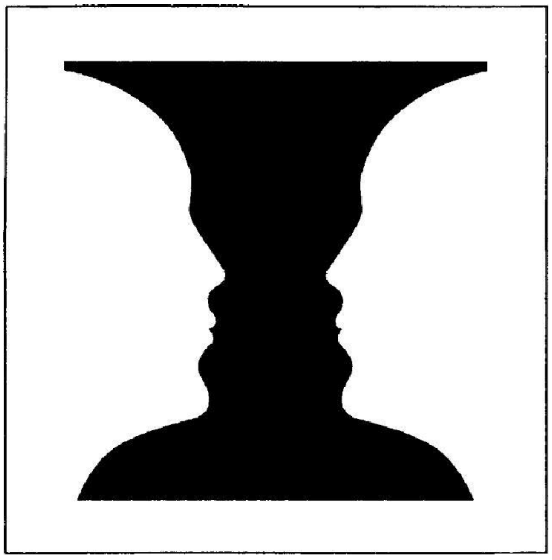


Figure 1.4. Rubin illusion, referred as "The Two Face, One Vase Illusion", where depending on the border assignment the content of the image swaps from a vase to two faces inf front of each other [91]

1.4.3 Border Ownership cells

In mammals, and some particular insects like Praying Mantis the perception of the three-dimensional space starts from the two-dimensional retinal images. To interact with the items in the scene, an agent needs to visually organise the perceived visual information acquired. The shape of an item is defined by the borders that allow one to distinguish it from the background. Gestalt psychologists were the first scientists to understand the importance of border ownership in perception, defining the need for an agent to distinguish the foreground from the background. Edgar Rubin proved how the detection of an item in the scene depends strongly on the border assignment process during the perceptual organisation of the visual information received [93]. The illusions proposed by Rubin (see Figure 1.4), where the content of the image changes depending on the border assignment, became a popular description for the figure-ground segmentation mechanism. There is no computational solution able to beat the visual system extracting features from the scene and solving the figure-ground task with the same robustness and reliability of the primate visual system [94], [95]. In 2000, Zhou et. al [92] found neurons in the visual cortex firing only if an edge belonged to an item as foreground (see Figure 1.5). The same cell mechanisms have been found in the human visual system [96] with strong evidence of dependence on attention mechanisms [97]. These cells are in the Secondary Visual Cortex (V2) with a significant connection to higher-level cortical areas [97]. Williford and von der Heydt [98] describe the Border Ownership (BO) cells coding reviewing the latest studies. These cells do not depend only on the local features detected in their receptive field but they strongly depend on the context. In recent studies, scientists tried to under-

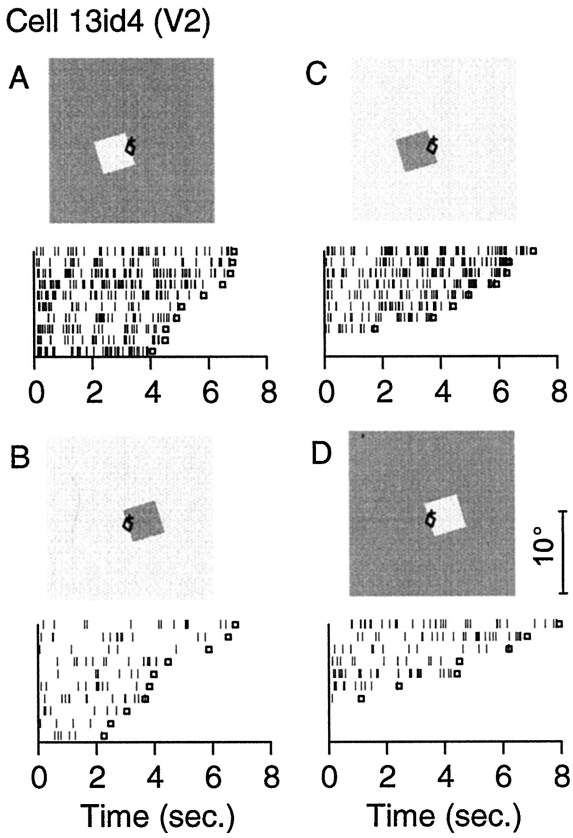


Figure 1.5. Representation of the firing rate of a Border Ownership cell in the Secondary Visual Cortex in monkeys [92]. The stimuli present a square as foreground and a background of a contrast colour (white and gray). The small black ellipse represents the orientation and the location of the receptive field. For each graph (A),B),C)&D) the raster plots show the response of the cell at the start of the fixation moment, where each row represent a different trial. The cell significantly responded when the edge belongs to the foreground on the left (A) and B)) and it poorly responded when the edge belonged to the square on the right (C) and D)). Figure adapted from Zhou et al. [92] under CCBY4.0.

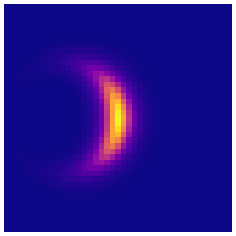


Figure 1.6. Representation of the Von Mises filter (0°) used in the Border Ownership layer to detect close contours.

stand how these cells modulate their action together with high-level cortex mechanisms proposing a hierarchical model with recurrence and lateral modulation respectively with the dorsal and the ventral stream [99]. The *dorsal stream* in the parietal lobe, also called ”where pathway” is the area of the cortex where objects are spatially located. The *ventral stream*, also called ”what pathway” corresponds to the temporal lobe areas, where an object is identified and recognised. In the work proposed by [99] they show evidence of BO dorsally-modulated simulated signal similar to the biological counterparts. The response is invariant to size, position and solid/outlined figures explaining the processing of contours in high-level areas of the cortex. The Border Ownership cells provide components of the perceptual organisation of the scene suggesting hypothetical grouping layers of cells using Gestalt laws to pool the information integrating the global contour of figures. These mechanisms could explain the higher saliency of areas containing possible objects, ”proto-objects”.

1.4.4 Proto-object Models

The concept of *proto-object* has been introduced as ”volatile units” of visual information that can be bound into a coherent and stable object when accessed by focused attention” [100], [101]. Early low-level processing defines whether a coherent structure, appearing in the retinal field, represents an object. This mechanism happens before the focus of attention [100]. The areas of the visual field that could potentially represent an object boost attention leading the agent’s gaze towards the proto-object. The same author who proposed the bioinspired saliency-based model [33] Christof Koch, proposed in 2006 together with Dirk Walther a biologically plausible system introducing the proto-object concept [102]. This model explains how an object can be detected even before being recognised within each feature such as intensity, colour and orientation. These features are then combined into single conspicuity maps and finally into a saliency map by a competitive mechanism. To estimate the proto-object region, feedback connections trace back to the most salient location at the conspicuity maps searching for the maps that contributed the most. The same three channels from [33], intensity, colour and orientation, extract features from the scene making a bioinspired model exploiting proto-objects to serialise object recognition in multi-object scenes [103].
During decades of studies, a proto-object model has been proposed for a humanoid robot to learn Gestalt laws [85]. The red, green and blue channels of each input RGB image are separated to generate colour opponency channels on which the edge extraction is performed. A watershed transformation is then applied to generate the proto-object upon which the saliency map is calculated. Yanulevskaya et al. [104] proposed a model assigning saliency

to the centre of objects rather than over edges using proto-objects extracted with image segmentation algorithms as coherent image regions. Proto-objects are exploited as coherent regions of neighbouring superpixels that share a common colour cluster also for clutter perception [105]. In 2014, Russell et al. [9] proposed a specific kernel to detect proto-objects following the Gestalt Laws using the same three channels of extraction (intensity, colour opponency and orientation). This model is a bottom-up biologically inspired architecture to generate a saliency map processing an RGB image as input. The model exploits Gabor filters to extract edges from the scene using an energy representation [106], [107] to emulate the response from contrast invariant complex cells. To obtain the response from centre surround cells the input image is filtered with a difference of Gaussians emulating the ON and OFF-center distinguishing light objects on dark backgrounds and dark objects on light backgrounds respectively. The processed outputs are then fed into the two main layers of the model: Border Ownership Pyramid and Grouping Pyramid. The Border Ownership Pyramid uses the Von Mises (VM) filter (see Figure 1.6) in different orientations to detect close contours from the scene. The information is then pooled together by the Grouping cells in the Grouping Pyramid layer giving saliency to the proto-object ROIs.
The output of the model is a saliency map of the scene, and the response is invariant to the size thanks to a classical computer vision pyramid mechanism [108]. This model is a baseline to generate the saliency map adding different channels of information to compete with each other for the representation of the scene. The work proposed by Russell et al. [9] has been extended adding different channels of information where a WTA mechanism selects the most salient region of the saliency map. In 2013, Molin et al. [87] proposed the same structure adding the non-directional motion information into the system. Hu [88] and Mancinelli [109] added depth perception, respectively using an RGB-D sensor and stereoscopic cameras with a number of known correspondence points. During the same period, Uejima et al. [89] added the texture information correcting boundaries through a bank of Gabor filters, but with a computationally expensive system. Despite these systems representing multimodal bottom-up saliency-based approaches, they do not exploit bioinspired and/or neuromorphic sensing or processing hardware to produce the saliency map. This project takes inspiration from them tailoring a fully bioinspired pipeline to exploit low latency for a robotic application requiring a fast reaction from the robot.

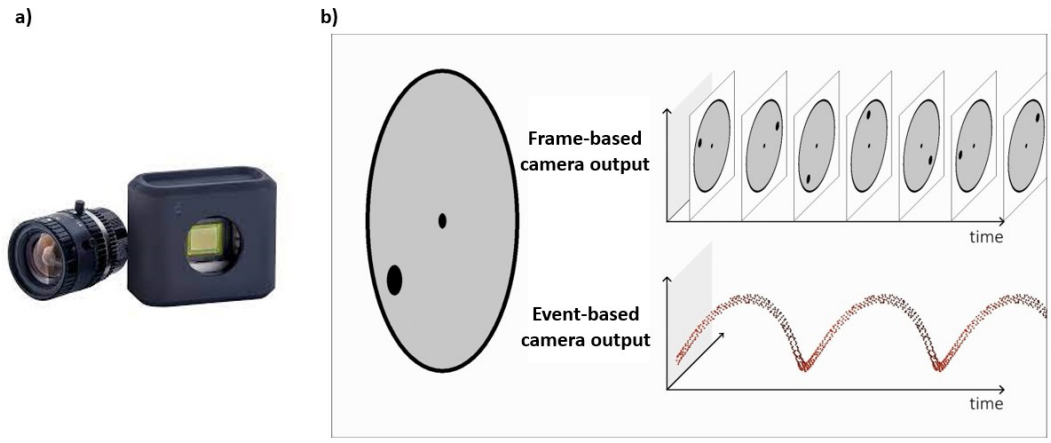


Figure 1.7. **a)** Event-driven Asynchronous Time-based Image Sensor (ATIS) camera. **b** Representation of the output from a classical frame-based sensor and an event-based sensor to a ball moving across the visual field. The image is adapted from [110].

1.5 Neuromorphic hardware

1.5.1 From the human retina to event-driven cameras

The human visual system is a complex organisation of highly structured mechanisms processing the light from the pupils and projecting the information to the retinas. The retina carries information from the photoreceptors (rods and cones) to the optic nerve. The majority of the cones are placed densely near the fovea, which allows light perception during the day. Rods instead, are sparse around the fovea, allowing us to perceive the presence of stimuli in the periphery of the visual field with low resolution, but with high sensitivity even in the darkness.

Upon detection of light, photoreceptors release glutamate to the bipolar cells, responsible for temporal contrast detection. The information goes then from the bipolar cells to the ganglion cells that generate spikes that travel from the retina to the brain. These 'vertical' connections work together with the Horizontal cells and the Amacrine cells ('horizontal' connections), which spatially integrate the information coming from the 'vertical' cells enhancing spatial contrast.

The combination of temporal and spatial integration of contrast changes and light decrements or increments encodes the visual signal associating and aggregating in time and space coherent features.

The human visual system allows seeing over a wide range of light intensities thanks to the *gain control* mechanism of adaptation, scaling the responses of the cells according to the ambient light level and shifting along different light intensities to obtain the same sensitivity [111], [112].
The human retina has been an inspiration throughout decades of studies inspiring scientists to build visual neuromorphic sensors loosely mimicking its functions. Lichtsteiner et al. [113] came out with a new generation of bioinspired event-driven cameras asynchronously encoding temporal contrast changes at pixel level at a high temporal resolution (see Figure 1.7). These cameras locally respond only where a brightness change occurs, providing a significant reduction in data, improvement of dynamic range and lower latency. The pixels of these cameras react to the contrast change based on a threshold mechanism, emitting an event (or spike) represented by the coordinates (x, y), the polarity (negative or positive change in contrast) p and the timestamp ts. The perception of darkness and light depends on the contrast of the stimulus and not on the absolute amount of light reflected. This exact mechanism happens in the human retina modulated by the gain control, while the contrast information is detected by the bipolar cells (ON-center and OFF-center-surround cells) through the lateral inhibition. Event-driven cameras have been used throughout the years exploiting their inherent characteristics for low and high-level vision algorithms [15] and are particularly suitable also for online robotic applications [114]–[118].
They have proven to be suitable for many applications such as motion estimation, depth and optical flow estimation [119], [120], angular velocity estimation [121] and contour motion estimation [122]. These cameras have been used also for neuromorphic and bioinspired algorithms for obstacle avoidance, detecting the motion direction [10], [123] or motion estimation [124], depth estimation solving the correspondence problem [11], [125] and many others.

1.5.2 Neuromorphic platforms

Neuromorphic computing platforms aim at capturing fundamental computational principles of neural systems. Decades of inspiration from nature have led to the studying of neuromorphic circuits inspired by biological principles [126]. Neuromorphic computing emulates with analog or digital circuits the nervous system mechanisms. The fundamental characteristics are listed below:

- The event-driven sensing is directly elicited by events in the sensory signal.
- The computation happens in the soma of each neuron of the population network.

- The digital communication of spikes guarantees robust transmission of information.
- The encoding of information exploits the spatio-temporal sequences of spikes, ”spike trains” from the network.
- Adaptation and on-chip learning. Feature adaption and learning at different temporal scales.
- In-memory computing, memory and computation are co-located.

Every neuron in the nervous system represents the fundamental unit of the brain, composed of an axon, dendrites and soma. Each neuron is able to communicate with neurons nearby via electrical signals, called *action potentials* based on the membrane potential changing over time. The classic membrane potential, at the resting state, is around -70 mV. If the membrane potential depolarises over a certain threshold the cell produces a spike. This mechanism represents the basis of the neuronal communication system allowing neurons to communicate with each other transmitting *spike trains* throughout the network.
Already in 1952, Hodgkin and Huxley [127] tried to model the ionic currents through the membrane as mathematical descriptions. After this important discovery, several neuron models have been proposed and studied, including the most used one: The Leaky Integrate and Fire (LIF) neuron [128].
In 1991, Mahowald and Douglas presented a circuit with the functional characteristic of a real nerve cell operating in real-time [129] imagining the creation of neurons on single chips. Over the years, from Mahowald and Douglas's HH silicon neuron, different platforms have been proposed with different characteristics and aims:

- **Analog**: Analog neuromorphic circuits depute the computation on physical circuits built to emulate a neuron following the characteristic differential equations. The global connections among neurons are digitally encoded to allow communication. BrainScaleS, BrainScaleS-2, DyNAPS, and ROLLS are the neuromorphic analog platforms currently available.
  BrainScaleS and BrainScaleS-2 allow fast simulations of large neural networks supporting neuroscience research and online learning. These platforms have accelerated computation thanks to the short time-constant to charge the capacitors.
  ROLLS and DyNAPS support the implementation of smaller networks in real-time as they feature the same time constants of biological neurons. ROLLS supports online learning, while DyNAPS can run pre-trained networks supporting temporal adaptation. These platforms have low flexibility due to the physical emulation of the neuron and low power consumption thanks to parallel in-memory computing.

- **Digital**: Digital neuromorphic platforms simulate the differential equations emulating neurons. This approach allows full flexibility to change the neuron model programming custom neurons behaviours. These platforms offer the possibility to create large networks using the cloud for deployment. SpiNNaker, Loihi and SPECK are the main examples, whilst SpiNNaker can offer boards with a different number of chips (different numbers of neurons to be created), Loihi is more compact and smaller, easily inserted via a USB input into the computer.
  SPECK is the world's first single-chip smart neuromorphic vision sensor based on DYNAP-CNN neuromorphic processor, combining a low-power SNN vision processor with an event-based sensor.
  These platforms are used for general-purpose computing of neuromorphic models.

Neuromorphic computing can be the key to more powerful computers. These platforms are extremely capable of reducing latency, promoting parallel computing as opposed to the Von Neumann architecture aiding serial computation due to the split of the memory from the computation. One important limitation of these platforms is the scalability of big networks with a high number of neurons.
The future of neuromorphic platforms is held in the new generation of neuromorphic hardware such as Loihi [130] or DYNAPs [131] that already proved to be suitable for address event representation (AER) [132], [133]. All of the platforms support AER. While the analog is a promising venue for final deployment, we prefer digital flexibility for model exploration. We used SpiNNaker which was available and fully functional at the time of my PhD.

1.6 Contribution of this work

Recent literature sees attention mechanisms exploited for several reasons such as classifying digits taking advantage of a spike-based approach but disregarding a biologically plausible pipeline [134], or gesture classification [135] through supervised learning.
The first model [134] integrates event-based visual and auditory signals exploiting a Spiking Neural Network (SNN). The two modalities are split into two separate sub-networks where the visual pathway sees several layers of convolution and pooling. The auditory modality exploits an architecture with several recurrent fully connected (FC) layers. The two modalities are fused by the attention-based cross-modality with an end-to-end training scheme for the overall multimodal network.
The latter one [135] describes the combination of bioinspired visual sensors producing events to be fed into a SNN using biology-grounded low-level computation. The archi-

tecture sees a feedforward pipeline with an attention neuron and three layers: the input, intermediate and output layers. The intermediate layer processes data only if the attention neuron is activated, namely the classifier processes only relevant data. This model classifies event videos of gestures where each neuron of the output layer learns a specific pattern and produces a spike when it is detected. Attention has also been used to judge the significance of event frames at the training stage [136] discarding irrelevant frames at the inference stage for different recognition tasks: gesture recognition, image classification, and spoken digit recognition. This model exploits a temporal-wise attention SNN looking for the correlation between event frames focusing on the most informative components of the input. All of these models do not take into consideration perceptual rules of perception such as the Gestalt laws, but most importantly they claim to focus on relevant parts of the scene without validating the results by comparing the saliency maps to the ground truth. Human fixational maps can be used to validate the outcome of the model by computing the similarity following standard analysis methods in the literature (Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji) & (AUC-Judd), Pearson's Correlation Coefficient (CC) and Similarity (SIM)) [35], [137]–[139]. These metrics judge several different aspects of the similarity [140] not allowing a single saliency map to perform well on all the metrics and determining how well the model approximates the eye fixations.

My idea of exploiting attention mechanisms for iCub is to create a building block for attention models starting from bottom-up approaches including Gestalt rules of visual perception to build up task-dependent mechanisms prioritising one channel at a time or a combination of channels. In doing so, the robot would be able to integrate mechanisms of alertness or selective attention depending on the scene dynamic. Moreover, the usage of biologically plausible pipelines on neuromorphic hardware allows full immersion in spike-based structures where redundant information is disregarded. The need for validation with fixational maps is necessary to understand how far the model reaches the ground truth where the task is not defined. Focusing attention is still a complex open topic where bottom-up and top-down cues are not well split and easily recognisable [141]. I started this project by designing three main channels of information for a basic event-driven saliency-based attention model: intensity, depth and motion. Intensity is inherently outputted from the event-driven cameras providing the information of contrast changes in the scene and the polarity associated. Depth is an important cue for attention mechanisms allowing a three-dimensional perception of items helping the interpretation of objects [45]. Eventually, motion plays a key role in being modulated by attention mechanisms during tracking and alertness [50], [51]. These three channels of attention represent to me the starting point towards a complex attention robotic schema.

The intensity and the depth channel directly fed into the proto-object model [9] allowing the perceptual organisation of the scene exploiting the Gestalt theories. Being an intrinsic alertness cue, the motion channel does not input into the proto-object model granting a response from the robot to anything approaching with a sustained speed.
The **Intensity channel** (see Chapter 2) has been implemented based on the RGB biologically plausible saliency-based model proposed by Russell et al. [9] modifying the input for the bioinspired sensors mounted on iCub. The first two layers of processing, edge extraction and detection of contrast, have been completely removed thanks to the inherent capabilities of the event-driven cameras to directly provide a similar scene representation. The model runs online using PyTorch on a GPU providing a saliency map as an outcome in 100ms. The model outputs an update of the saliency map every time events occur. The biological pipeline has proven to suit the event-based representation for online robotic applications showing low latency and reduced power consumption removing layers of processing. The system is able to detect proto-objects in different scenarios slightly removing clutter from the scene. Moreover, the saliency map is provided also with dynamic scenarios and with fast motion (2000 [px/ms]). The limitation of this implementation is the lack of the robustness of the system to maintain attention focused on a specific target. This system does not take full advantage of the low latency and low power consumption of a fully neuromorphic pipeline on a neuromorphic dedicated platform. In this respect, the intensity channel has been further investigated by implementing a fully spiking-based version (see Chapter 3) of the event-based model using the neuromorphic platform SpiNNaker. We changed dramatically the structure of the model to suit a SNN pipeline exploiting at best the event-based characteristics of the bio-inspired algorithm. The SNN architecture exploits populations of proto-object neurons where each neuron encodes the output of one VM kernel. The model has different orientations of the VM filter and different sizes ensuring scale invariance to the system. We further benchmarked the output validating the model with the ground truth and analysing the qualitative and quantitative results. The model shows a great capacity in removing the clutter from the scene, significantly improving the outcome with respect to the PyTorch implementation. This implementation proved also a great capability in reducing the latency to 4ms to produce a saliency spike output, confirming the relevance of a fully spiking-based pipeline for applications where the latency of perceptual processing is crucial, such as in robots that need to interact smoothly with the dynamical environment. Although these significant improvements, taking inputs from the full resolution of the ATIS cameras implies more than one physical SpiNNaker board to run the model. A real robotic application setup requiring three physical SpiNNaker boards would be a challenge if the robot would need to be free to move in the future. Diversely, the simplification of the system reducing the visual field

to fit the model in one board would require a field of view of 50x50 pixels. This approach would significantly diminish the visibility of the scene affecting the saliency results. The use of neuromorphic platforms in computer vision is still an open problem due to the high number of neurons required to cover the full visual field.

The **Depth channel** (see Chapter 4) has been introduced thanks to the asynchronous event-based bio-inspired cooperative matching algorithm proposed by Firouzi et al. [11] solving the correspondence problem in a multiple objects scenario. The disparity extractor is providing an online asynchronous disparity map of the scene which has been directly fed into the proto-object model [9]. The disparity extractor has been implemented using C++ feeding the PyTorch attention network. The entire system extracts the online disparity map and detects proto-objects providing a saliency map in 170ms. In this work, we benchmarked the response of the model comparing the outcome with real ground truth fixation maps using saliency-based metrics known in the literature. The system proved the capability to provide an online response for the robot feeding into the proto-object model a robust disparity map of the scene. Furthermore, the model significantly improved in stability avoiding the focus of attention to jump from one proto-object to another one. The event-driven depth attention model focuses on the closest target as it is the most probable to be reached by the robot. The model was able to select the closest proto-object discarding non-proto-object items generating more events, also prioritising the proto-object detection over closer non-proto-objects.

The **Motion channel** (see Chapter 5) is implemented using SpiNNaker. This model relies on the Spiking Elementary Motion Detector (sEMD) proposed by Milde et al. [10], where the consecutive activation of two neighbouring receptive fields in the visual field is used to measure motion in a given direction. The implementation of this motion detector builds up from the biologically inspired models proposed by Hassenstein and Reichardt [142] and Barlow and Levick [143] exploiting the time-to-travel method [144]. The fundamental of this system lies in each elementary motion detection unit being selective to motion only in one cardinal direction, suppressing motion in the anti-preferred direction[1].

The model has been modified to exploit the eccentric structure of the human retina, seeing the size of the receptive fields increasing going from the fovea to the periphery.

My contribution to this model is the implementation of an ”eccentric” down-sampling replacing the uniform initial down-sampling of the original model. This structure allowed a wider speed range detection of moving objects reducing the overall Mean Firing Rate (MFR) needed to detect motion direction. The model is yet not robust for complex scenarios opening a crucial problem for the removal of ego-motion in dynamic scenes. Moreover,

[1]Time Difference Encoding (TDE) neurons are sensitive to a preferred motion direction. LR (Left to Right), RL(Right to Left), TB(Top to Bottom), and BT(Bottom to Top). The anti-preferred direction is opposite to the preferred one (i.e. Right to Left or Left to Right)

the retina structure can be further exploited creating a log-polar representation with angular directions obtaining the angle direction of an incoming entity without making any additional computation.
These three channels represent the work done towards a biologically plausible event-driven saliency-based attention model on bioinspired sensors and neuromorphic platforms.

Chapter 2

Proto-object based saliency for event-driven cameras

This work has been published.

Iacono, M., D'Angelo, G., Glover, A., Tikhanoff, V., Niebur, E., & Bartolozzi, C. (2019, November). Proto-object based saliency for event-driven cameras. In 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS) (pp. 805-812). IEEE.

2.1 Personal Contribution

Theoretical Implementation	Code development	Experiments design	Experiments execution	Analysis of the results	Paper writing	Academic authorship
yes	no	yes	yes	yes	yes	Second author

2.2 Authors

Massimiliano Iacono, Corresponding First Author, Istituto Italiano di Tecnologia
Giulia D'Angelo, Istituto Italiano di Tecnologia - The University of Manchester
Arren Glover, Istituto Italiano di Tecnologia
Vadim Tikhanoff, Istituto Italiano di Tecnologia
Ernst Niebur, Johns Hopkins University
Chiara Bartolozzi, Istituto Italiano di Tecnologia

2.3 Authors Contribution

C.B. and M.I conceived the main idea behind the work with the help of A.G and G.D. M.I. developed the code for the event-based proto-object model. M.I. and G.D. designed the experiments with supervision from C.B. and A.G. M.I. and G.D. conducted experiments. M.I. and G.D. analysed the experimental results. M.I. and G.D. wrote the manuscript with the supervision of C.B, A.G., V.T and E.N. G.D. and V.T. made the supplementary video accompanying this work.

2.4 Abstract

Autonomous robots can rely on attention mechanisms to explore complex scenes and select salient stimuli relevant for behaviour. Stimulus selection should be fast to efficiently allocate available (and limited) computational resources to process in detail a subset of the otherwise overwhelmingly large sensory input. The amount of processing required is a product of the amount of data sampled by a robot's sensors; while a standard RGB camera produces a fixed amount of data for every pixel of the sensor, an *event-camera* produces data only for where there is a contrast change in the field of view, and does so with a lower latency. In this paper, we describe the implementation of a state-of-the-art bottom-up attention model, based on structuring the visual scene in terms of proto-objects. As an event-camera encodes different visual information compared to frame-based cameras, the original algorithm must be adapted and modified. We find that the event-camera's inherent detection of edges removes the need for some early stages of processing in the model. We describe the modifications, compare the event-driven algorithm to the original, and validate the potential for use on the iCub humanoid robot.

Multimedia Material

`Video:https://zenodo.org/record/7112769`

2.5 Introduction

Attentional selection is crucial for biological organisms to react to the most important stimulus at any given time, like a fast-moving predator from which they must immediately escape, or a single red apple amongst green foliage. Fast selection is paramount to

real-time interaction (and survival) of the system in a dynamic world and enables detailed further processing of a small region of the visual input since all of the input cannot be fully processed by the brain in real time [145], [146]. Autonomous robots can similarly take advantage of attention mechanisms to reduce the computational load for visual processing when confronted with the vast amount of information in the world, and choose the most appropriate behaviour. As in all resource-restricted systems, in which it is impossible to fully process all sensor information simultaneously, choosing the most important sensory signals is important. Attention solves this problem by subdividing visual input into segments, identifying the most relevant of these segments, and processing them sequentially in the order of decreasing relevance. This process is formalised in the concept of the saliency map [147].

To be behaviourally relevant in robotic applications, the computation of the saliency map and the selection of relevant stimuli have to be performed in the shortest possible time, while minimising the use of computational resources. In a previous study [44], it has been demonstrated that using event-cameras [113] meets these targets, substantially reducing the latency and the computational cost of the feature-based saliency model initially proposed in [75]. In that model, the saliency of a stimulus is defined by features such as intensity, colour, orientation, etc. However, recent work based on the concept of proto-objects proved to better explain perceptual saliency [9]. Proto-objects are regions of the scene that *potentially* correspond to physical objects. This concept of "perceptual organisation" of visual scenes derives from the work of early Gestalt psychologists that developed "Gestalt laws", e.g. continuation, proximity, and convexity [83]. Russel et. al. [9] implemented a subset of such laws into a simple computational model.

Such an approach is relevant for robots that operate within a human environment, as it is behaviourally relevant to, e.g., quickly locate potential objects upon which the robot can act. The algorithm [9] is divided into three main stages: 1. a centre-surround filter enhances the contrast of the stimulus and acts as pre-processing for the extraction of edges, 2. border ownership cells represent edges and, importantly, signal in their firing rates the location of foreground objects relative to the location of their receptive fields, as observed in primate extrastriate cortex [92], [98], and 3. grouping cells which respond to regions enclosed by several borders, thereby representing the presence of a proto-object [98].

In this work, we adapt the proto-object-based saliency model from [9] to use the visual signal provided by an event-camera [5] mounted on the neuromorphic iCub humanoid robot [1], to allow the robot to quickly select object candidates and pass only the relevant regions of interest (ROI) to further modules to perform the high-resolution processing required for recognition, grasping and manipulation.

Event-cameras have more recently gained interest for robotics applications. They have been designed to capture the stimulus-driven activation of retinal cells, meaning that each pixel responds individually (rather than as part of a whole image frame), and only to *change* in the light falling on this pixel. The result is a low-latency, asynchronous visual signal that describes the edges and contrast change of objects in motion. For robotics, this means lower processing requirements and a faster response [114], [148]. The proto-object approach to saliency, in which an object is defined by borders which are bound together at the grouping cell stage, is particularly suited for using event-camera signals, as it only responds to the edges and outlines of objects. However, as event-cameras perform a different computation at the silicon level, they produce a different output and visual representation than traditional frame-based systems. For this reason, it is necessary to adapt existing models, developed using input from frame-based cameras, to work with a new data representation and encoding. Interestingly, we found that, as the event-camera more closely corresponds to a biological vision system, the processing layers in the original proto-object model designed to respond to contrast changes (i.e. centre surround cells [149]) are no longer required in the event-driven model.

In this paper we describe the differences in the visual signal of the event-camera and how the original proto-object algorithm can be adapted. We design and tune the algorithm for our intended robotic application: an iCub robot quickly identifying potential objects. The event-driven algorithm is then validated on identical stimuli used in the original study, and also on new stimuli typical for the iCub environment.

2.6 Proto-object based saliency

In this section we describe the adaption of the proto-object algorithm to the event-camera data, which is formed by three different layers of processing: centre-surround, border ownership, and grouping cells.

The baseline algorithm, from which we form our event-driven version, and to which results are compared, is described in [9]. Attention models such as this have been inspired by [33] in which images are split into different feature maps, separately providing information about intensity, colour and orientation. The feature maps are then filtered by means of centre surround kernels, inspired by the organisation of visual cortex [150]. Final stages of the model aim to integrate and further process the feature maps to provide the final saliency map.

The main difference of [9] compared to similar studies is the exploitation of the proto-

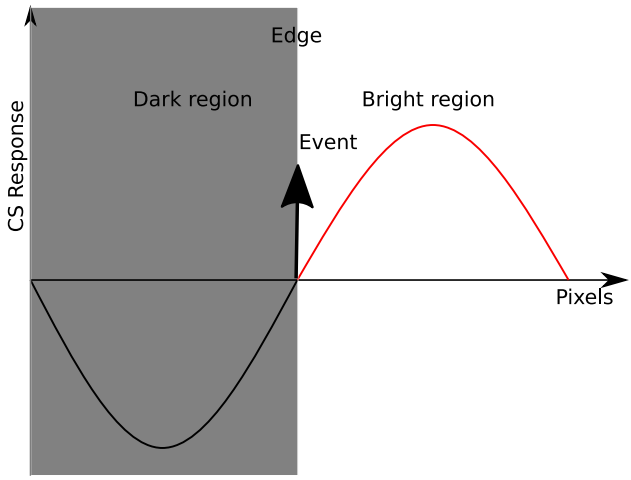


Figure 2.1. CS filters with ON (red) and OFF (black) centres respond to the light and dark side of the edge respectively. An event-camera will instead produce a an event directly on the edge location, with a polarity dependent on the edge's direction of motion.

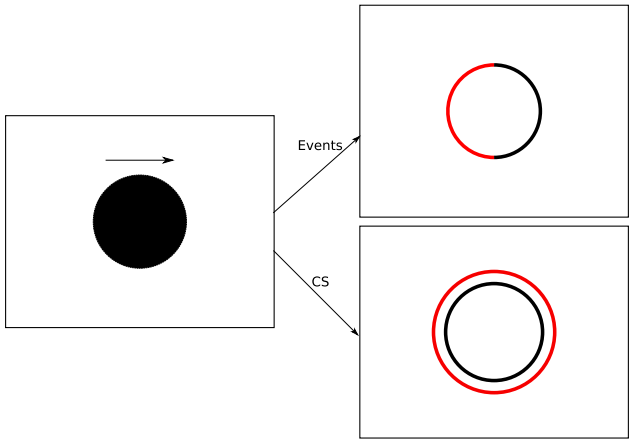


Figure 2.2. As an object moves, an event-camera produces events of opposite polarities on either side of the object. Processing an RGB image with CS filters instead results in negative (black) and positive (red) polarities on the inside and the outside of the object.

object concept [101], where (partially) closed contours that might correspond to objects provoke a strong activation on the saliency map. The final processing layers are divided in two main stages: border ownership and grouping. The first aims to respond to individual edges that potentially form a (generally convex) border of an object. The second groups the elements of these potential borders, and regions enclosed by several object borders generate high response of grouping cells.

The peak response of the initial centre-surround (CS) filter is offset with respect to the edges that cause the cell's excitation. The result is that the filter response does not occur on the actual edges of objects. Figure 2.1 shows qualitatively this behavior in the presence of a high contrast region, in relation to the spike (event) generated by the event camera. Figure 2.2 additionally illustrates the different locations at which positive and negative

polarities occur. At the border-ownership stage, the output of CS cells is convolved with kernels generated from the von Mises distribution. As shown below (Fig. 2.5), the filters have an offset which moves the saliency signal back to the edge location. The filter orientation also has a preferred direction, eliciting a stronger response in the presence of edges whose convexity matches one of the filters. This is important for segmentation, because, according to Gestalt principles [83], objects tend to be convex around their centres. Border ownership cells are computed from the output of Complex cells, composed of of odd and even Gabor filters, making responses to edges phase-invariant.

The response of border ownership cells is then integrated at the grouping cells stage, which is responsive to multiple borders that show proximity and continuity features [83]. The grouping cells also receive an inhibitory signal from filters with the opposing preferred side. This mutual inhibition suppresses the response from isolated edges, which aren't particularly convex in either direction. Border ownership is a mutually exclusive property: a given edge either belongs to one object or another, and potential ambiguities are resolved in the human visual system in the form of perceptual rivalry [151], an effect that has been shown to be implemented in primate visual cortex [152].

For more details on the original implementation we direct the reader to [9]; the present study adapts this algorithm to suit the visual signals of an event-camera.

2.6.1 Event-camera

The pixels of an event-driven cameras only elicit a spike when they detect a change in the light intensity. As each pixel fires independently of all others, the resulting output is an asynchronous stream of events carrying information about the location, the time and the polarity (light to dark or dark to light) of such changes. The events have low latency, on the order of a few microseconds, allowing for gap-free tracing of moving edges on the image plane.

In our implementation, each event (or spike) v is represented by its coordinates, polarity and time stamp, $v(x, y, p, ts)$. The events received from the camera are accumulated into a binary matrix Vas well as two other binary maps V_+ and V_-, encoding for positive and negative polarities respectively. All maps have the same size as the sensor, which in our case is 304×240. We consider a fixed number of events in a time window W at each iteration, but multiple events occurring at the same location are considered as a single one according to

$$V_+(x, y) = \begin{cases} 1, & \text{if } \exists\, v \in W \mid p_v(x, y) = 1 \\ 0, & \text{otherwise} \end{cases} \tag{2.1}$$

where $p_v(x, y)$ is the polarity of the event occurring at coordinates (x, y) that can be either 1 (positive) or 0 (negative). Similarly we fill $V_-(x, y)$ with events where $p_v(x, y) = 0$. The resulting maps are then sent to the border ownership computation where we select the borders which likely belong to objects.

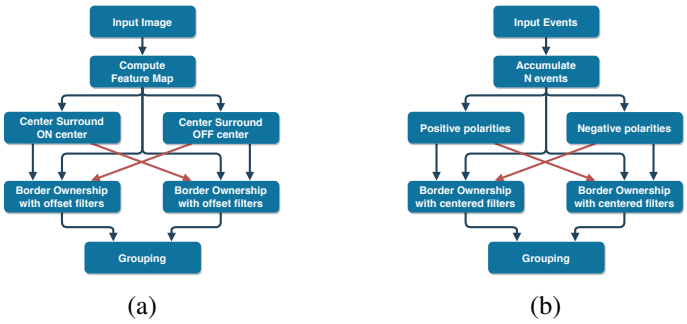
As the event-camera does not have colour sensitivity, the colour-based feature map was not used in our implementation. In addition, for initial algorithm simplicity, the orientation filter was also removed. The implication is that our algorithm will be equally responsive to edges independent of their angle. In comparison, an orientation filter map would instead make a single horizontal line more salient amongst many vertical lines. We made this simplification since we assume that its effect on the saliency computation is minor. If necessary, an orientation filter can be reintroduced in future work.

2.6.2 Center-Surround

The pixels of event-driven cameras produce an asynchronous stream of events every time they detect an illumination change, making them natural contrast and edge detectors. Assuming a dark object moving on a light background, as shown in Fig. 2.2, processing RGB camera inputs through CS filtering of opposite polarities would produce negative responses on the inside of the object and positive outside. In the event camera, we obtain "positive" (off-to-on) spikes on the leading edge of the object (the side towards which the object is moving) and negative spikes (on-to-off) at its trailing edge. Even though the information that we receive from these two types of representations is very similar, the interpretation of the scene comes from different processes.

Furthermore, as reported in the original paper [9], Russell et al. used a center-surround mechanism of both polarities detecting a light object on a dark background and vice-versa, i.e., ON-center and OFF-center receptive field. However, as the events occur only on edges, implicitly containing the polarity information, we can make the assumption that the information that would normally come from the CS can be inherently found in the output of these sensors.

Not only can we exploit this information to save some computation steps of the algorithm, but we can also assume that the spiking pixels already give us the precise location of the edges in the image plane. We therefore remove the CS layer from the event-driven implementation. Figure 2.3 highlights the differences between the algorithm from [9] and ours.



(a) (b)

Figure 2.3. Comparison between border ownership algorithm in [9] (a) and ours (b). The red lines are inhibitory signals. In our implementation the Center Surround is not used as the polarities of the events already encode a similar information. Also we don't compute feature maps, but we use the events as they come from the sensor. For clarity we don't show multiple feature maps in (a).

2.6.3 Border Ownership

Despite using the same filters as in [9], modifications were made due to our application to robotics. In short, filters were made more rounded and less responsive to straight edges. The border ownership cells are modelled by the Von Mises (VM) distribution:

$$VM_\theta(x, y) = \frac{\exp(\rho \cdot R_0 \cdot \cos(atan2(-Y, X) - \theta)}{I_0(\sqrt{X^2 + Y^2 - R_0})} \tag{2.2}$$

where X and Y are the kernel coordinates with origin in the centre of the filter, R_0 is the radius of the filter, θ its orientation and I_0 is the modified Bessel Function of the first kind. We added the ρ parameter which determines the arc length of active pixels in the kernel, allowing to change the convexity of the kernel. For values $\rho < 1$ the filter becomes more sensitive to convexities rather than straight lines, making it more suitable for the proto-object detection task. Suitable tuning of this parameter improves the robustness of the model, making it more suitable to detect curved lines. We have empirically found that a value of $\rho = 0.2$ is good for detecting convexities while rejecting straight lines at the same time. The orientations we use are $\theta = [0, 45, 90, 135]$.[1] Additionally, since we have already an "in-place" response on edges due to the event-camera, as shown in Fig. 2.2, the filter response does not need to be moved back to the edge locations. At this stage we instead use filters which are centered on the position of the peak activity. To do so we apply a simple translation to eq. (2.2) as follows:

$$\begin{aligned} X' &= X + R_0 \cos(\theta) \\ Y' &= Y + R_0 \sin(\theta) \end{aligned} \tag{2.3}$$

[1]The opposite orientations with opposite polarity will be used to compute Equation 2.4.

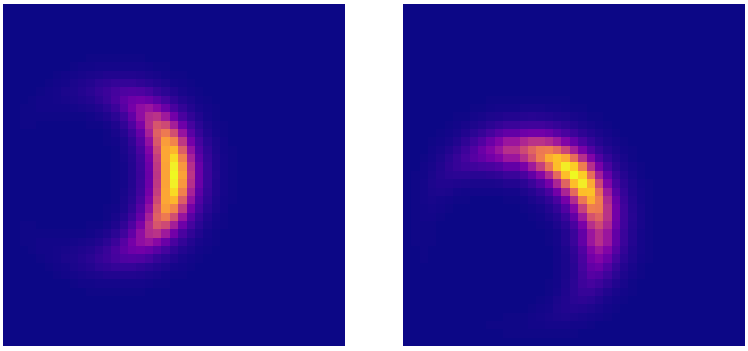


Figure 2.4. Example Von Mises (VM) filters used at the border ownership stage at 0 and 45 degrees. The centre of the filter is at the peak of the filter response, i.e. these filters are 'in-place'.

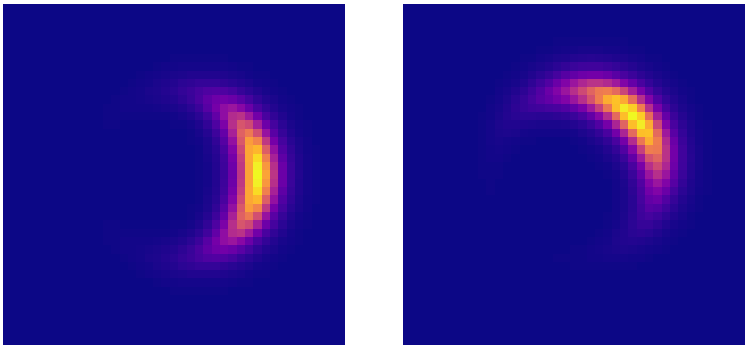


Figure 2.5. Example Von Mises (VM) filters used at grouping stage at 0 and 45 degrees. The centre of the filter is offset from the peak response.

The resulting filters are shown in Fig. 2.4. The final border ownership response is then computed as follows:

$$\begin{aligned} B1_\theta &= V \odot (\lfloor V_+ * VM_\theta - wV_- * VM_{\theta+\pi} \rfloor \\ &\quad + \lfloor V_- * VM_\theta - wV_+ * VM_{\theta+\pi} \rfloor) \\ B2_\theta &= V \odot (\lfloor V_+ * VM_{\theta+\pi} - wV_- * VM_\theta \rfloor \\ &\quad + \lfloor V_- * VM_{\theta+\pi} - wV_+ * VM_\theta \rfloor) \end{aligned} \tag{2.4}$$

where $\lfloor \cdot \rfloor$ is a linear rectification operation, $*$ a convolution, and $\odot$ an element-wise multiplication operator. The factor w weights the inhibition between competing polarities and orientations. The higher its value the harder it is for one orientation to dominate over its opposite, filtering out ambiguities. In other words, the border ownership cells are excited by the presence of a convex edge at a certain orientation θ and inhibited by activity of the same edge with opposite polarity and orientation. As a result, the borders that show only one preferred side and only one polarity are preserved, and all the others get inhibited. This mechanism helps reject clutter and noise because responses from both polarities are suppressed, as are straight lines. The rectification ensures that no inhibitory signal gets propagated to later stages of the computation. Finally the element-wise multiplication by

V masks the response of the border ownership located only where events exist. The result of the border ownership computation gives each event a score based on its likelihood of being a border regardless of its polarity according to [92]. The resulting $B1_\theta$ and $B2_\theta$ encode the dominant orientations in the range $0 \leq \theta < \pi$ and $\pi \leq \theta < 2\pi$. The response of the border ownership layer is then passed to the grouping cells.

2.6.4 Grouping Cells

The activity of border ownership signal cells is grouped by the grouping cells (G). The grouping mechanism moves the energy towards the centre of the objects, from multiple boundaries, and thereby enhances proximity and continuity patterns [153]. Grouping is achieved using the same kernels as in Eq. 2.2, and the standard (no translated) filters are used, in which the centre of the filter does not coincide with the maximal response of that filter.

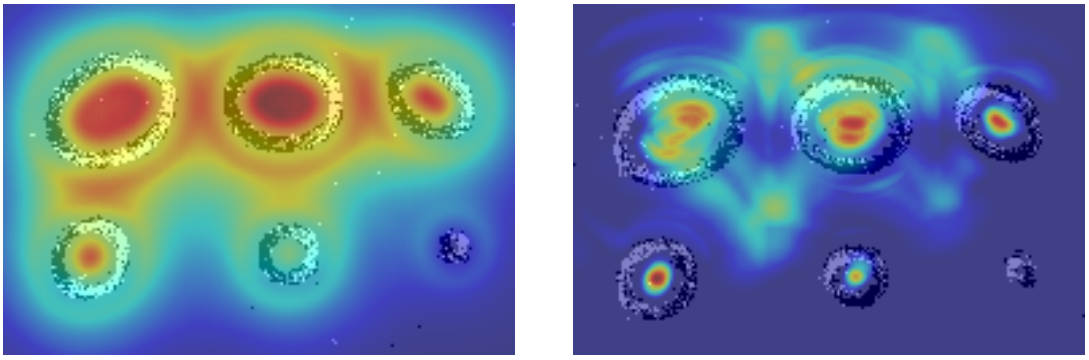
The excitatory component of the grouping cells is computed as:

$$\begin{aligned} G1 &= \sum_\theta B1_\theta * VM_\theta \\ G2 &= \sum_\theta B2_\theta * VM_{\theta+\pi} \end{aligned} \tag{2.5}$$

$B1$ and $B2$ are highly responsive to opposite convexity, this is why in Eq. 2.5 we apply two opposite VM filters. The aim of this is to move all the response coming from the object edges to the centre of the object. This process would also affect the object's surround, which can lead to ambiguity in-between objects. We introduce an inhibition mechanism for the grouping cells that reduces inter-object interference and preserves the saliency inside the object. First we compute the inhibitory signal as shown in Eq. 2.6. Different from Eq. 2.5, we apply kernels of the opposite orientation, because we are suppressing the activity of the non-preferred side of the border ownership cells:

$$\begin{aligned} G1^* &= \sum_\theta B1_\theta * VM_{\theta+\pi} \\ G2^* &= \sum_\theta B2_\theta * VM_\theta \end{aligned} \tag{2.6}$$

We find the maximum value within the map and subtract it from from the remaining elements, eq.(2.7). The reason is that peaks of activity can be found inside the objects, where responses from all orientations overlap, and by subtracting the maximum value these peaks get suppressed. If we then take the absolute value of the resulting map, we



(a) (b)

Figure 2.6. Results of the calibration (a) without grouping inhibition and (b) with grouping inhibition. With correct inhibition calibration, high peaks of response can be found in the objects of sizes 20-30-40-50 pixels, which is the desired sensitivity for a typical humanoid robot application.

are left with an inhibitory signal which is mostly concentrated outside and in-between the objects. Fig. 2.6 shows the effects of the inhibition on the final result.

$$\begin{aligned} G1^* &= |G1^* - max(G1^*)| \\ G2^* &= |G2^* - max(G2^*)| \end{aligned} \tag{2.7}$$

The final grouping is computed as follows:

$$G = (G1 - G1^*) + (G2 - G2^*) \tag{2.8}$$

2.6.5 Scale invariance

All the computation steps mentioned so far are performed at several different scales to obtain object size invariance/tolerance. To achieve this, the feature maps are arranged into a pyramid, in which each level is down-scaled by a $\sqrt{2}$ factor. To obtain the saliency map we collapse all the levels into one by applying the normalisation method described in [33].

2.7 Validation and experimental results

We implemented the algorithm to work on the neuromorphic iCub humanoid robot [2], equipped with a pair of ATIS [5] sensors coupled with a pair of traditional cameras that we use for validation of the algorithm, both located in the robot's eyes. The two cameras share the same field of view and are calibrated to have pixel to pixel correspondence [118]. The ATIS camera has 304×240 pixels, whereas the RGB camera has a nominal resolution of

1920×1080 but for our experiments the images are down-scaled to 320×240. The algorithm is implemented in C++ and runs online on the robot. In the following experiments, the iCub looks at a number of either static or dynamic objects. Since static objects do not elicit any response from the event-cameras, the eyes are programmed to move in small circles introducing relative motion between them and the cameras. The circular motion has been chosen to span all possible orientations and capture all the edges in the scene.

2.7.1 Calibration

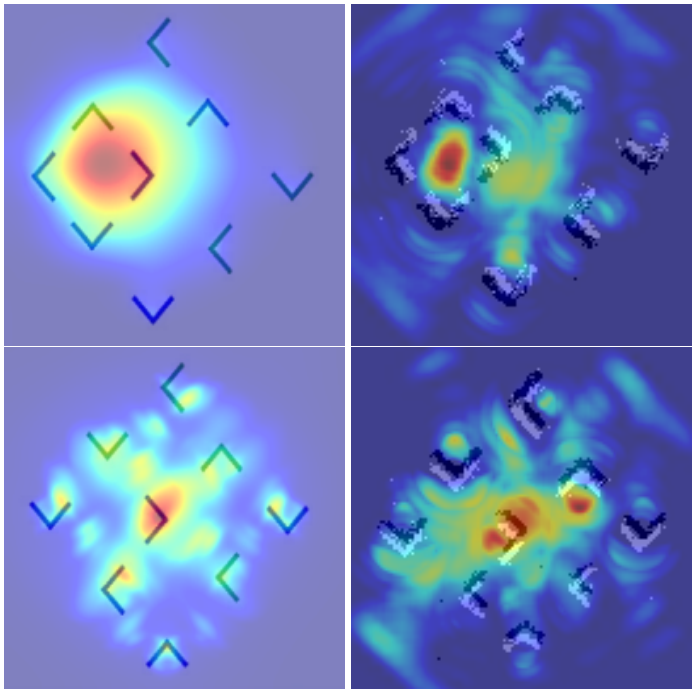
In a first set of preparatory tests, we tuned some parameters of the cells in the model to match the correct range of object sizes and positions that are relevant for the robot. Specifically, we target applications where the robot can grasp and manipulate objects, tailoring the model to optimally respond to objects whose size in the image plane is between 25 and 50 pixels. We chose this range of sizes according to the typical object that the robot can interact with, which is constrained by the robot workspace and grasping capabilities [154], [155]. We ran a calibration test in a controlled scenario to set the right filter size for best response to the situation of interest. For this purpose, we put in front of the robot's eyes a calibration pattern with six circles of radii ranging from 10 to 60 mm. In this condition, we set the pyramid depth to 1, in order to find the smallest desired size by increasing the VM filter radius up to the point where we obtain a high peak in the middle of the object. Once calibrated for the smallest size, we increase the number of pyramid levels until the algorithm responds to the largest desired object size. Results of the calibration process are shown in Fig. 2.6. With this empirical procedure, we set $R_0 = 10$ (see Equation 2.2) and 5 levels of pyramid. Table 2.1 shows the values used for each parameter of the model.

2.7.2 Comparison with original algorithm

We carried out a series of experiments to benchmark our model against the original work [9]. To this aim, we printed a set of images, on which we computed the saliency map using [9], and showed them to the event-driven cameras mounted on the robot. Fig. 2.7 compares the response of both models to two pattern of corners. In the first pattern (top), four

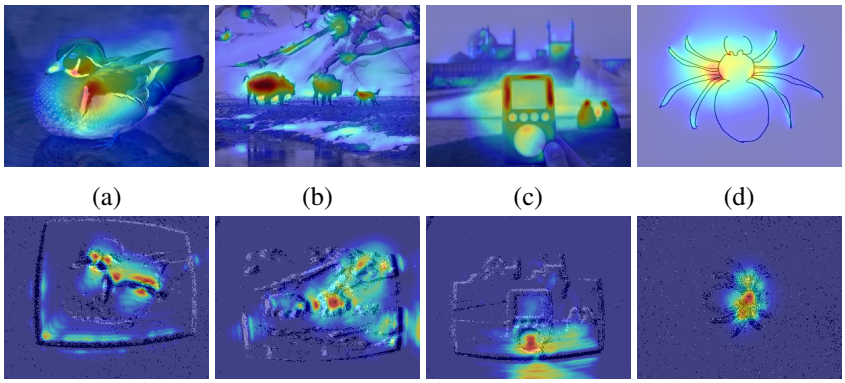
Parameter	Value
R_0	10
ρ	0.2
ω	3
Pyramid levels	5
Orientations	0, 45, 90, 135

Table 2.1. Values of the parameters used in the experiments



(a) RGB camera (b) event-camera

Figure 2.7. The response to proto-objects for a closed stimulus (top) and to a similar stimulus but without the enclosed shape (bottom), comparing the original RGB implementation to the proposed event-driven implementation.



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

(e) (f) (g) (h)

Figure 2.8. Comparison of saliency maps from the grouping cells response obtained with the original algorithm (first row) and our implementation (second row). The four pictures come from a dataset for saliency benchmarking [139]

corners enclose a squared area; these are generally perceived as parts of the edge of an object, while the second pattern does not contain a similar "proto-object" [156], [157]. Both models correctly select the corners in the second pattern and show a peak activation of the saliency map in correspondence of the space enclosed by the four corners corresponding to the proto-object (left side of the image plane). In contrast, random orientation of corners (bottom) do not generate the perception of an object, and both models reflect this by gen-

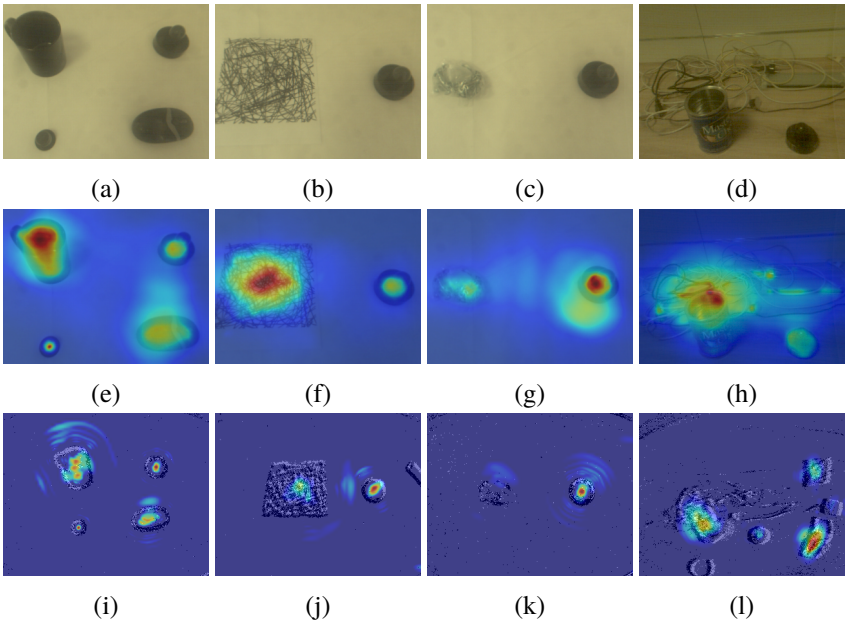


Figure 2.9. Saliency response for different stimuli including multiple objects, clutter texture and cluttered objects. The top row shows the original image, the middle row shows the response of the original RGB algorithm and the bottom row shows the response of the proposed event-driven algorithm.

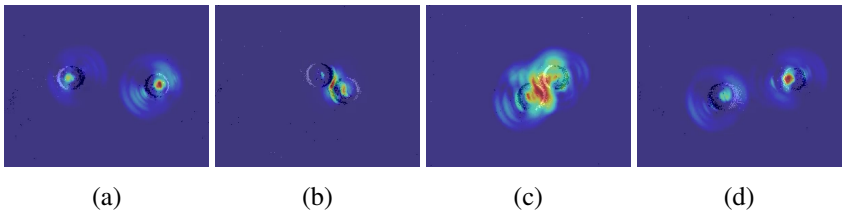


Figure 2.10. Effect of distance between objects. From left to right, the objects are first far apart (a), only when the objects share a contact point they are perceived as a single object (b,c). They are again detected as distinct ones with increasing distance (d).

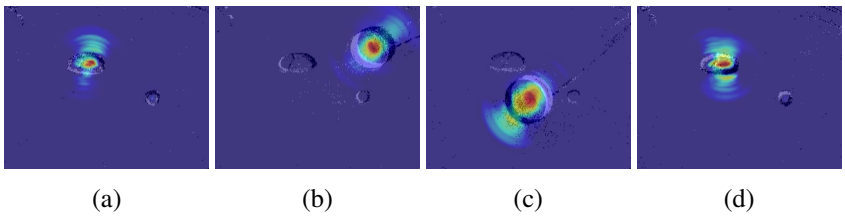


Figure 2.11. Saliency map for static and dynamic stimuli: (a) Two static objects are placed on the table and the focus of attention is localised on the left (bigger) object. (b, c) As the dynamic stimulus enters the field of view, it captures the robot's attention. When the moving object gets out of sight (d), attention goes back to the static object.

erating weaker, disorganized activation patterns. Fig. 2.8 compares the output of the two models to some images taken from the saliency benchmark dataset CAT2000 [139]. All Figures present a coherent and comparable response except for one, Figs 2.8c/2.8g. This might be due to the enhancement in curvature of the VM filter we adopted that generates higher response in the presence of circular shapes like the one in the picture.

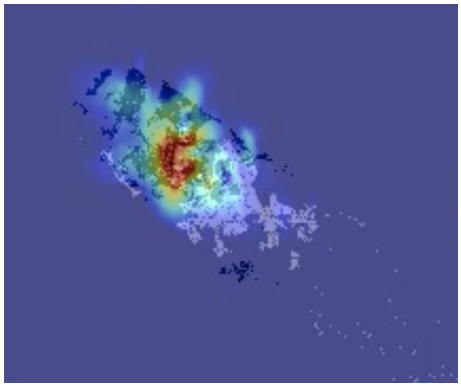


Figure 2.12. Saliency map of rapidly moving object

Fig. 2.9 shows the algorithm behavior in a realistic scenario, with objects of different sizes and shapes as well as distractors (in form of cluttered cables) placed in front of the robot. In this case, we recorded simultaneously RGB images and events [118], and run the original model on the RGB image. The output of both models is consistent, confirming that the adapted algorithm responds to the presence of proto-objects in the scene.

While the output of the proposed event-based and original models are qualitatively comparable, they produce different responses to the input images in Figures 2.9b and 2.9c, where the textured objects on the left side is less salient in the event-driven model. The strong inhibition factor $w = 3$ might explain this result. Moreover, the added inhibition mechanism explained in Section 2.6.4 would be useful to suppress the saliency of regions with events of both polarities, that often correspond to noise, clutter, or flicker stimuli.

2.7.3 Moving objects

Finally, we tested the event-driven model with dynamic scenes. Fig. 2.10 shows some snapshots of a sequence where two objects roll on a desk and collide. The model shows two peaks when the objects are far apart. When the two objects are closer, the saliency map shows interference between the two objects, generating a single peak.

The role of inhibition in the grouping layer is crucial to suppress the activity elicited by edges with opposing curvatures, i.e. contours of two different objects which would increase saliency in the space between two objects. This mechanism works as long as the two objects are not too close: when two approaching objects are touching each other, the algorithm is not able to distinguish them anymore. The grouping mechanism reduces the representation of objects to their simplest possible form. This is in agreement with the Gestalt law of proximity, where “objects or shapes that are close to one another appear to form groups”. Algorithmically, the collision between the two objects generates a large

number of events which causes attention to get focused on that point. Fig. 2.11 shows snapshots of a sequence with the robot observing at static objects when a new object enters the field of view. In Fig. 2.11a, the static objects generate peaks in the saliency map. In Figs. 2.11b and 2.11c the dynamic stimulus captures the robot's attention as soon as it gets in the field of view of the camera. When the dynamic object leaves the field of view, the attention goes back on the previous object, Fig. 2.11d. An advantage of the event camera is that it can process very fast motion because it is not limited by frame rates. To demonstrate that this advantage translates into saliency processing of rapidly moving objects, we tossed a ball in front of the cameras which moved with a speed of $\approx$2000 px/s. Fig. 2.12 shows that our method successfully places attention on the ball.

2.8 Conclusion

The aim of this work is to adapt a proto-object attention model [9] to work with the neuromorphic event-driven cameras embedded on the iCub humanoid platform. The overarching goal of this approach is to endow the robot with a low-latency, computationally efficient attention system that we believe is fundamental in an image processing pipeline for a robot which has to act in a dynamic and unconstrained environment. As event-driven cameras encode the visual signal in a radically different way with respect to frame-based cameras, it was necessary to tailor the model and correctly interpret the sensor output and the effect of the different filters on the novel input. Specifically, the event-driven sensor acts as an edge extractor, functionally replacing the first layer of the frame-based model, based on centre-surround filters. However, further modifications to the border ownership and grouping layers are required to correctly process the output of the event-driven camera. Specifically, we separated on and off events in two parallel streams and modified the inhibition connectivity pattern in the grouping layer to take into account the difference between polarities in the leading and trailing edges of objects.

We carried out preliminary qualitative experiments to prove the consistency of our implementation with the theoretical model and test its limits. In general, the implementation of the proto-object-based saliency model proposed in this paper produces an output that is consistent with the original model. We tested the model with static objects and images, but also extended the scenario to dynamic scenes, with moving objects. In the event-based representation, saliency is related to the speed of the object since increased speed increases the number of events of the faster object within the scene. Event cameras naturally produce a bias towards this type of stimulus as the moving object generates the most events for the camera.

The C++ implementation of the event-driven model runs online on the neuromorphic iCub and is capable of selecting objects in a range of sizes that are typically used when the robot performs grasping and manipulation tasks. The use of event-driven cameras – that efficiently compress the signal and performs part of the computation on a chip – leads to a computationally efficient implementation. However, the proposed implementation is based on a hybrid solution, where the events are accumulated in frames that are then convolved with Von Mises filters in the border ownership and grouping cell layers. While this implementation is certainly helpful in characterising the algorithm and proof that the results are comparable to the original model, a fully spiking implementation of the model will further increase efficiency and latency: Events elicited by the sensor travel asynchronously along the hierarchy, the computation is restricted only to the filters that receive input events and, as soon as there is enough activity in a region, the network can produce a result that can be almost simultaneous with the stimulus presentation. The spiking implementation of the proposed model and quantitative analysis of its performance in terms of attentional selection, efficiency and latency are the goals of current development. Additionally, we are exploring the possibilities of learning the kernels by exploiting gradient-based techniques.

2.9 Reflections & Conclusions

The first attempt to modify a biologically plausible model of attention mechanisms, by adapting of an RGB saliency-based proto-object model to receive an event-based input has been advantageous for several reasons.

The event-based implementation needed a modification of the Border Ownership computation (Equation 2.4) subtracting the contribution of kernels of the opposite orientation and polarity. This is due to the leading edge polarity of an object being opposite to the trailing edge. The computation results in detecting curve edges of θ angles and their opposites with competing polarities. Final $G1^*$ and $G2^*$ represent the activity of the non-preferred side to be suppressed. The final Grouping stage sum up the contribution of opposite angle kernels finally detecting the proto-object.

Due to the event-driven camera properties, the stereotyped eye movements are needed for static scenes emulating biological microsaccades [158]. The choice of the parameter in Table 2.1 refers to empirical values found for: R_0, ρ and ω. As shown by the calibration results in Figure 2.6 (b), the 5 Pyramid levels cover the real-world objects' size range explored.

The frame-based implementation runs on MATLAB requiring minutes of execution due to the nested loops in the code. This implementation is not meant to run online for a robotic

application. The event-based proto-object model optimised the code in C++ for an online application and it has been further implemented in Python exploiting PyTorch to obtain a saliency map every 100ms. Thanks to the intrinsic edge-extraction mechanism performed by the event camera, we could get rid of the first two layers of processing (contrast detection and edge extraction) promoting a reduced pipeline with the consequent further reduction in processing loads.

The time to obtain a saliency map is fairly similar to the time needed to compute a human saccade, demonstrating a comparable and ”natural” reaction in the time of the system. Although the model clearly detects proto-objects, the comparison presented in Figure 2.8 shows a non-equal response between the models. The same response can be seen also in Figure 2.9. The evProto shows overall a more local response and is less centre-biased. The frame-based model indeed has a further centre-biasing mechanism which has not been implemented. The model response can still be defined as coherent because the salient points are found to belong to the same items selected as most salient from the fame-based implementation. Furthermore, the response in Figure 2.11 suggests a probable propensity of the model to be driven by the number of events. This model well represents the initial intensity channel for my Thesis project proving good performances in proto-object detection in different situations. The static and dynamic experiments, as well as the benchmark dataset, show a reliable response from the event-based proto-object model.

Despite all of the above-mentioned reasons, the pipeline can not be considered fully neuromorphic due to the use of the event-frames representation used as input to the model. The system is still not robust when placed in an environment with multiple objects with similar salience, where the selection among them becomes jittery. Objects have similar salience if the selection is based only on contrast. The jittery response seeing the salient point ping poing among the items in the scene has been addressed in Chapter 4 prioritising the selection of the closest proto-object. A mechanism of WTA with hysteresis and self-excitation could further solve the problem allowing for a robust fixation over a specific target. As an alternative solution, a feature channel disambiguating the selection and boosting the salience over a particular proto-object could be added.

This work clearly shows qualitative results detecting objects in the scene still not entirely removing the clutter. Even though these results show a good response from the model in different circumstances they do not validate the model outcome. The analysis of the response must be investigated quantitatively definitively assessing the saliency map in comparison to the ground truth. A quantitative analysis has been further added in the next Chapters (3 and 4).

Chapter 3

Event driven bio-inspired attentive system for the iCub humanoid robot on SpiNNaker

This work has been published.

D'Angelo, G., Perrett, A., Iacono, M., Furber, S., & Bartolozzi, C. (2022). Event driven bio-inspired attentive system for the iCub humanoid robot on SpiNNaker. Neuromorphic Computing and Engineering, 2(2), 024008.

3.1 Personal Contribution

Theoretical Implementation	Code development	Experiments design	Experiments execution	Analysis of the results	Paper writing	Academic authorship
yes	yes	yes	yes	yes	yes	First co-author

3.2 Authors

Giulia D'Angelo, Corresponding First Co-Author, Istituto Italiano di Tecnologia - The University of Manchester
Adam Perrett, First Co-Author, The University of Manchester
Massimiliano Iacono, Istituto Italiano di Tecnologia
Steve Furber, The University of Manchester

Chiara Bartolozzi, Istituto Italiano di Tecnologia

3.3 Authors Contribution

G.D. conceived the main idea behind the work with the supervision of C.B. G.D. developed the theory for the spiking-based proto-object model with the help from A.P. A.P. developed the code on the neuromorphic platform with the supervision of G.D. G.D. designed the experiments. G.D. and A.P. conducted experiments. G.D. analysed the experimental results. G.D. and A.P. wrote the manuscript with the supervision of C.B. and S.F. M.I. gave valuable feedback and helped edit the manuscript.

3.4 Abstract

Attention leads the gaze of the observer towards interesting items, allowing a detailed analysis only for selected regions of a scene. A robot can take advantage of the perceptual organisation of the features in the scene to guide its attention to better understand its environment. Current bottom-up attention models work with standard RGB cameras requiring a significant amount of time to detect the most salient item in a frame-based fashion. Event-driven cameras are an innovative technology to asynchronously detect contrast changes in the scene with a high temporal resolution and low latency. We propose a new neuromorphic pipeline exploiting the asynchronous output of the event-driven cameras to generate saliency maps of the scene. In an attempt to further decrease the latency, the neuromorphic attention model is implemented in a spiking neural network on SpiNNaker, a dedicated neuromorphic platform. The proposed implementation has been compared with its bio-inspired GPU counterpart, and it has been benchmarked against ground truth fixational maps. The system successfully detects items in the scene, producing saliency maps comparable with the GPU implementation. The asynchronous pipeline achieves an average of 16 ms latency to produce a usable saliency map.

3.5 Introduction

Visual attention guides the perception of the environment [159]. It is a mechanism that selects relevant parts of the scene to sequentially allocate the limited available computational resources to smaller regions of the field of view. In the animal world, this is coupled with eye movements, aimed to sequentially centre the selected region within the

highest resolution region of the retina [160]. The detailed analysis only of salient regions of the visual field can dramatically reduce the computational load of processing the full visual field at once. In a similar manner, a robot working in real-time can exploit visual attention advantageously to optimise the use of computational resources. The motivation of this work is to produce an analogous reduction in computational loads for autonomous systems. Robots, such as the humanoid robot iCub [1], need to generate fast and precise response to autonomously interact with the environment reacting to external stimuli. Recent studies in computer vision have exploited the concept of attention for different tasks: classifying MNIST handwritten numbers only on regions of interest (ROIs) of the visual field with the 1.07% error [60], fixation prediction adding audio cues [161], visual search [58], and object recognition, where it has been demonstrated that attentional selection (based on saliency) increases the number of regions where objects are identified with random ROI selection [59].

Attention has attracted interest since the first psychological experiments where Yarbus et al. [25] were recording the fixation points of subjects examining different pictures. Since then, attention has been modelled in order to understand its underlying neural implementation, and to equip artificial agents with similar capability to obtain a reasonable perception of the scene [75]. Attention is a complex mechanism that results from the interplay of a bottom-up process that is driven by the physical characteristic of the stimuli and top-down effects that depends on priors and goals [141]. Diverse studies tried to model the bottom-up components of attention. Some proposed the use of the saliency map formalism [19], [28], [162]. A saliency map is the representation of visual saliency in a scene, where each item appears to be interesting (salient) based on the observer visual exploration [147].

Specifically, selective attention extracts features from the environment and explains the situation as fast as possible filtering what is not necessary to understand the scene. [163]. The widely used feature-based saliency model [75] extracts in parallel multiple different visual features and finds regions of high contrast within each feature channel. Their contribution defines the saliency of each point in the field of view. The weight of each feature map can be modulated to model the effect of top-down mechanisms competing with each other for the representation of the scene. This model was then augmented [102], by integrating principles of perceptual grouping of individual components that reflect “Gestalt laws” as proximity, common fate, good continuity and closure [83].

These principles give perceptual saliency to regions of the visual field that can be perceived as ”proto-objects” [146], [145].

A proto-object describes regions of the visual field that may coincide to real objects in the physical world, referring to the human ability to organise part of the retina stimuli into structures [164]. The work of Russell et al. [9] improved [102] by creating a filter capa-

ble of detecting partial contours. Recent studies added other sources of information to the proto-object model such as motion [87], depth [88] and texture [89]. Further, a new line of research has started to develop these types of models using event-driven cameras as input. In these cameras, the contrast change in the scene is outputted asynchronously, with high temporal resolution, low latency, and most importantly, reducing data rate. For a real-time application in a robotics scenario this leads to a faster response given the low processing required [114], [148].

Adams et. al. [165] exploited the address-event representation (AER) and the neuromorphic platform SpiNNaker to allow the humanoid robot iCub [1] to perform real-world tasks fixating attention upon a selected stimulus. Rea et. al. [44] exploited visual attention for a bio-inspired pipeline using event-driven cameras (ATIS cameras) [5] mounted on iCub, the neuromorphic robot [2]. This implementation [44] exploits the low latency of the event cameras, further increasing the speed of the response towards online attention, but does not include the proto-object concept, that was later included by modifying a frame-based proto-object model [9] in a way that is suitable for event-based cameras [166]. The implementation proposed by Iacono et al. [166] adapts the proto-object model based on RGB cameras to event-driven input, using the contrast feature maps naturally encoded by event-driven cameras. However that work didn't fully exploit the advantages given by the sensor. In fact events were accumulated over time generating frames that were then processed using a GPU. In an attempt to decrease latency and computational cost we implemented the model proposed in [166] on the SpiNNaker neuromorphic computing platform [167], that is able to properly exploit the asynchronous output of the event-based cameras. SpiNNaker is a dedicated neuromorphic computational device which provides a digital platform to model spiking neural networks at large scale in real time. Using an asynchronous and highly parallel architecture, large numbers of small data packets can be processed, which in most applications represents spikes being sent between biological neurons. This provides an ideal computational tool for event based processing.

The platform supports asynchronous spiking models that propagate events from the sensors in the network. Such models yield minimum processing latency, most of which depends on the propagation across layers and on the accumulation of sufficient information [168]. The contribution of this work is the validation of the model implemented on SpiNNaker (SNNevProto) through a direct comparison with the event-driven proto-object (PyTevProto) (i.e. its counterpart implemented on GPU using PyTorch). We compared the two models using the dataset from [166] (SalMapIROS) and benchmarked both against ground truth fixation maps [169]. We analyse the trade off between accuracy, number of neurons, computational cost and latency.

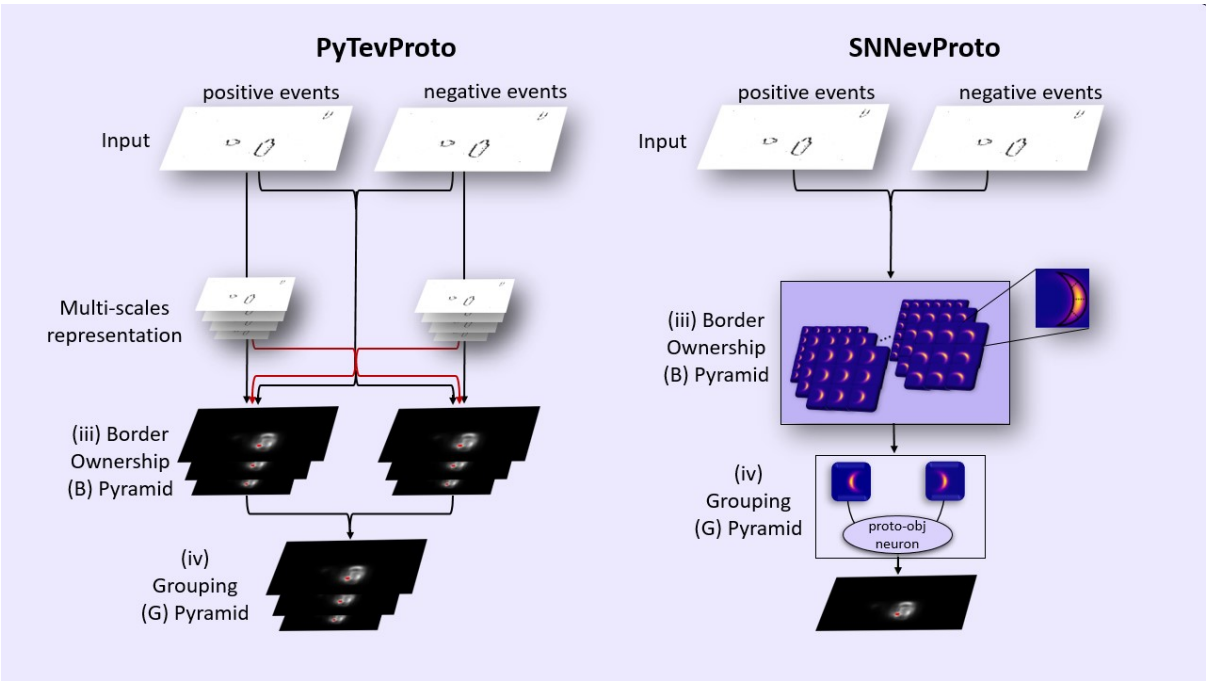


Figure 3.1. An overview of the model architectures for the PyTevProto (on the left) and the SNNevProto (on the right). The events are split based on the polarity and fed into the two models as input. The event-based model generates different scales by subsampling the "event-frame" and creating a pyramid. The resulting scaled "event-frames" are convolved with VM filters at 4 different orientations (Border Ownership Pyramid) and grouped at the Grouping Layer directly processes the input with the two layers o Border Ownership and Grouping Pyramids. The red lines are inhibitory signals. The spike-based implementation processes the events asynchronously exploiting layers of VM shaped neurons at different scales and rotations. The Proto-Object Neurons (Grouping Pyramid Layer) integrate the response connecting VM filters with opposite side and pool the response from different scales. The outcome of both models is the saliency map.

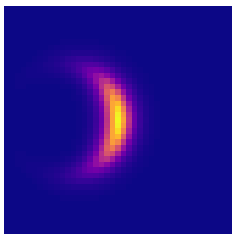


Figure 3.2. Representation of the VM filter described in Eq. 3.1 at 0 °

3.6 Event-based Spiking Neural Network proto-object saliency model

This work takes inspiration from the bio-inspired saliency-based proto-object model for frame-based cameras initially proposed by Russell et al. [9] and its event-camera adaptation [166]. The former is composed of three channels: intensity, colour opponency and orientation, competing with each other to represent the scene. Its core is composed of four layers: Center Surround Pyramids (CSP), Edge Pyramids, Border Ownership and the Grouping Pyramid (see Fig. 3.1).

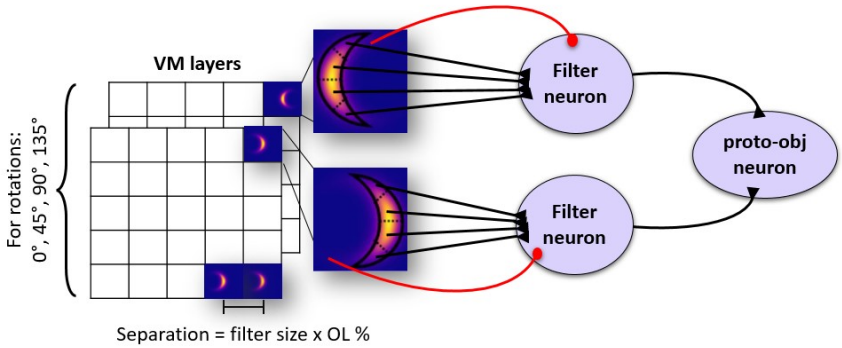


Figure 3.3. Representation of a VM layer and its connections. Each VM filter is split in 4 sections all connected to the same Filter neuron. The area around the "active" part of the neuron (moon shaped yellow region) is connected to the Filter neuron with an Inhibitory connection (red lines). This stage of the model represents the Border Ownership pyramids detecting close contours. Two complementary VM filters with opposite orientation are then connected to the same Proto-Object Neuron (Grouping Pyramid) to identify possible proto-objects. This structure is repeated for each layer with different orientations of the filter: 0°, 45°, 90° and 135°.

The CSP layer convolves the input image with a difference of Gaussians kernel to detect regions in the scene with either positive or negative contrast, emulating the Center Surround (or Bipolar) cells present in the retina [149], [170]. In parallel, the system convolves the RGB image with Gabor filters, emulating the edge extraction done by the Primary Visual Cortex [171]. The Border Ownership and Grouping Pyramid implement the "Gestalt laws" of continuity and figure-ground segmentation, mimicking the neurons in the Secondary Visual Cortex area, which are mostly selective to edges [92]. All the computation steps are performed at several scales to obtain object size invariance/tolerance. In the Border Ownership layer the output of the CSP is convolved with curved Von Mises (VM) filter (see Fig. 3.2). The convolution with four different orientations of the filter detects partial contours of objects. All filters in the same location are connected via inhibitory connections to each other creating local competition for the dominant orientation. The output is then pooled by the Grouping Pyramid which combines oppositely rotated contours oriented to the same centre forming a partially closed contour. Closed contour activity is captured by the proto-object neurons whose combined activity creates the saliency map. In [166] we have adapted this model to run using the output of event-driven cameras. Here we take a step further, implementing the model with spiking neurons on neuromorphic hardware.

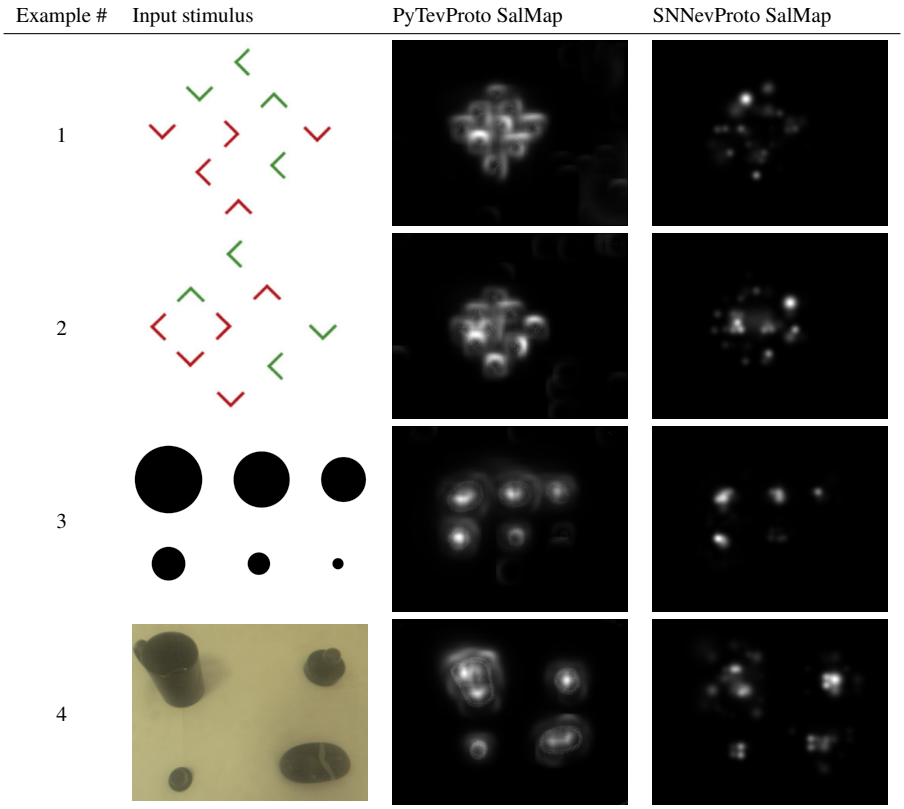


Figure 3.4. Qualitative comparison among the PyTevProto and the SNNevProto. From the left column to the right column: the example number, a RGB image representing the scene shown to iCub (the input stimulus), PyTevProto saliency map and SNNevProto saliency map. These examples are a selection from 13 scenarios of the SalMapIROS dataset. The events are recorded directly from the event-driven cameras mounted on iCub's eyes. The objects and the 2D printed patterns are placed on a desk in front of the robot. The RGB input images are only for a better visualisation of the input stimulus.

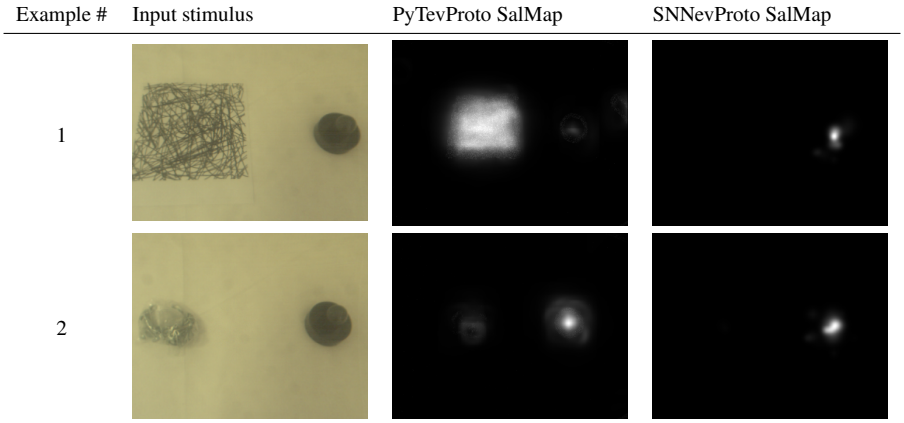
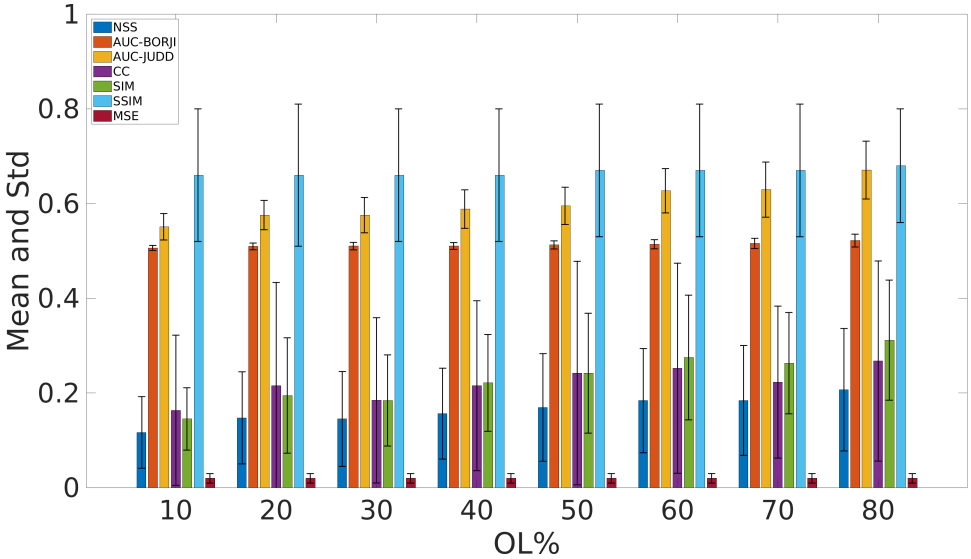
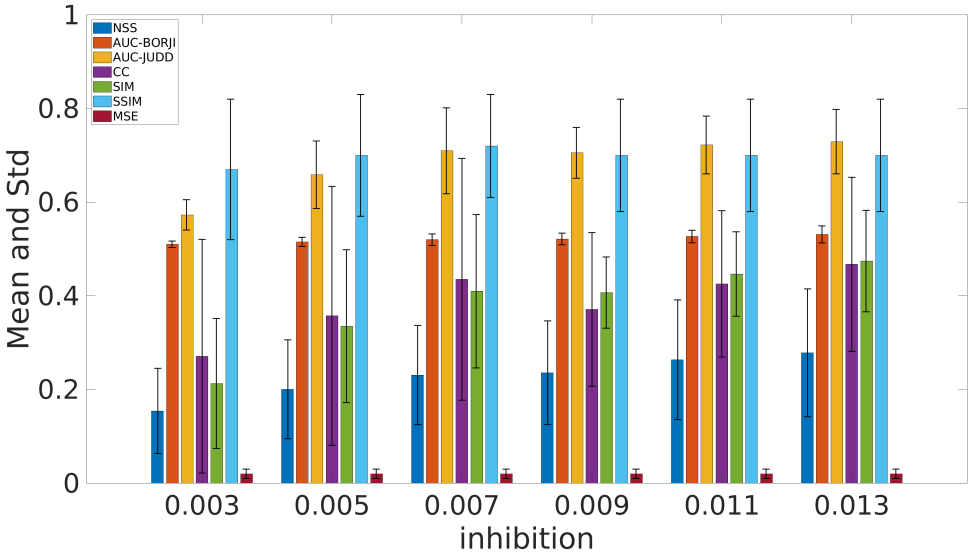


Figure 3.5. Qualitative comparison among the PyTevProto and the SNNevProto. From the left column to the right column: the example number, a RGB image representing the scene shown to iCub (the input stimulus), PyTevProto saliency map and SNNevProto saliency map. This table show only results from clutter experiments of the SalMapIROS dataset. The events are recorded directly from the event-driven cameras mounted on iCub's eyes. The objects and the 2D printed patterns are placed on a desk in front of the robot. The RGB input images are only for a better visualisation of the input stimulus.



(a)



(b)

Figure 3.6. Comparison with different metrics evaluating the similarity between the the SNNevProto saliency maps and the PyTevProto saliency maps [166] using the SalMapIROS Fig. 3.4) exploring different OL percentages (**a**) exploring a range of inhibitions (**b**) (μS conductances) with fixed OL percentage at 60%. The metrics used are: the Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji) & (AUC-Judd), Pearson's Correlation Coefficient (CC) and Similarity (SIM) [35], [137]–[139], Structural Similarity (SSIM) and Mean Square Error (MSE). A higher score is better for all excluding the MSE where the lower score determines similarity.

Event-driven camera's pixels asynchronously produce an event every time a local illumination change occurs providing the information of positive or negative change in contrast. As such, they perform an inherent operation of edge extraction that can functionally be equivalent to the edge extraction performed by *center-surround* (CS) cells in the frame-

OL%	# of neurons	# of SpiNNaker boards
10%	10428	3
20%	12000	3
30%	15801	3
40%	22266	3
50%	30306	6
60%	48878	6
70%	82084	12
80%	176248	24

Table 3.1. Table showing the number of neurons and SpiNNaker boards required given a percentage of overlapping for the VM filters. The spalloc server was used to run these jobs which allocates boards in multiples of 3.

based model. A similar contrast change information is provided by the CS cells [172]. The event-driven camera does not obtain the local contrast change due to lateral inhibition as in the CS cells, but rather due to the relative motion between the camera and the scene. The two processes are different but the related outcome, the edge extraction and the contrast information, are similar. These inherent capabilities can be used as substitutes for the first two layers of processing in the event-based version of the saliency-based model [166]: *center-surround* filtering and edge extraction. In fact, assuming a dynamic scene where a dark object is moving over a white background the leading edge would produce negative events and the trailing edge positive events, therefore providing information about the object contrast with respect to the background. In the PyTevProto model implementation running on GPU, the output from the event-based cameras is used to create frames of events divided into positive and negative polarity. The frames of events are fed into the Border Ownership layer following the process explained above.

This work proposes a new fully spiking based pipeline, with dedicated neuromorphic hardware, aiming to improve the speed and reduce the latency of the model. The SpiNNaker neuromorphic platform [167] acts as a computation medium modelling the SNN in a feedforward architecture (see Fig. 3.1). The neural model mimics the cells as populations of current-based leaky integrate and fire neurons.

These neurons process the data coming from the ATIS cameras in form of events carrying the information of the position in the visual field, polarity (positive or negative contrast change) and the timestamp of the event. The VM filter, shown in Figure 3.2, is a kernel designed to respond to curved edges that can potentially delimit a closed area. They are formalised as a curve (Eq. 3.1) with the largest value at its midpoint providing the ideal

input->segment	segment->filter	input->filter	filter->proto-object
0.02%	0.8%	0.0013%	0.75%

Table 3.2. The percentage firing thresholds for different population connections, input->filter is the only inhibitory connection. Percentage firing threshold is the percentage of the pre-synaptic population that need to fire to produce a spike in the post-synaptic population. Inhibitory connections do not induce a spike but are scaled in the same fashion. This metric is used to standardise weights across varying convolutional kernel sizes.

shape to respond to closed contours:

$$VM_\theta(x,y) = \frac{\exp(\rho \cdot R_0 \cdot \cos(atan2(-y,x) - \theta)}{I_0(\sqrt{x^2+y^2-R_0})} \tag{3.1}$$

Where x and y are the kernel coordinates with origin in the centre of the filter, R_0 is the radius of the filter, ρ determines the arc length of active pixels in the kernel allowing to change the convexity of the kernel, θ its orientation and I_0 is the modified Bessel Function of the first kind. The VM output is then thresholded to reduce sensitivity to localised activity:

$$e(x,y) = \begin{cases} 1 & \text{for } VM_\theta(x,y) > 0.75 \\ -1 & \text{else} \end{cases} \tag{3.2}$$

Where $e(x,y)$ describes whether the pixel at (x, y) is connected to the filter neuron with excitatory synapses ($e(x,y) = 1$) or inhibitory synapses ($e(x,y) = -1$)(see Fig. 3.3). Connection weights, w, are determined using Eq. 3.3 where n is the size of the pre-synaptic population and p is the percentage firing threshold for that particular projection between populations. A value of $5\mu S$ is chosen as it is the minimum weight at which one excitatory input spike produces a spike in the post-synaptic neuron in this implementation of conductance-based neurons. Inhibitory connections are scaled using the same method but do not produce a post-synaptic spike. Values of the percentage firing thresholds of connection weights can be found in Table 3.2. The filters are used as convolutional kernels which are tiled over the whole image[1].

$$w = \frac{5}{pn} \tag{3.3}$$

This implementation of the model is a Spiking Neural Network where the first layer is covered with VM filters spaced with strides relative to their size [2]. Consequently, each

[1]The percentage of overlapping between filters is a parameter of the system.
[2]The distance between two centres of the VM filters depends on the overlapping percentage and consequently on their size.

VM filter has its own receptive field onto the input layer. Therefore each incoming event triggers a specific pixel belonging to a specific receptive field in the field of view. Each VM filter is composed of four rotationally distributed segments. As the inputs are discrete spikes generated by an event-based camera it is possible for noise and other artefacts to produce a high number of events in a small area unrelated to the visual scene. Splitting the VM filter into four sections helps to reduce the sensitivity to localised activity, aiding the filter to respond more selectively to input spikes arranged in the shape of the VM. As the strides of the convolutional kernels are relatively large, appropriate control of VM filter activity is important to reduce undesired spikes and, therefore, inaccurate saliency map generation. Each filter segment is connected to a neuron representing the entire VM filter. The refractory periods of the segment neurons and input weights to the filter neuron are balanced to require all segments to fire within a narrow temporal window to produce a spike. In addition, all spikes within the filter region that are not part of the VM kernel will have an inhibitory contribution to the combined filter neuron, effectively increasing the selectivity to the VM shape (see Fig.3.2). The grouping cells, called proto-object neurons, pool the output of VM complementary cells that form a close contour representing proto-objects (see Fig. 3.3). The output of the convolution, and the subsequent output of the proto-objects which form the saliency map, are all represented as spikes emitted by a neuron. The filters exist in 4 rotation pairs with their complementary filters rotated 180°, evenly distributed from 0-135°, and in 5 spatial scales (104, 73, 51, 36, 25 pixels^2). Over each layer the VM filters are placed overlapped with each other. Overlap is related to stride used in the convolutional layers of neural networks. Instead of measuring how much the filter has shifted relative to the previous it measures how much it is overlapping with the previous. The overlap among the VM filters is important to define the robustness of the model. In biology, cell receptive fields are often overlapped for robustness, ensuring a response even if a cell no longer functions [173], [174]. Over time, cells overlapping have been used as a way to avoid the aliasing problem in bio-inspired models [175]. The overlapping percentage (OL) increases resolution and accuracy and it is directly linked to the number of neurons required in the implementation and, hence, its power and computational cost (see Table 3.1). We therefore decided to use the OL as a parameter of the model to be explored. A percentage is used to ensure a uniform overlap at multiple spatial scales.

Each VM filter is connected with its mirrored one (VM in Fig. 3.3) of the opposite side creating a sub-population. All projections between sub-populations share a common weight as described in Eq. 3.3. This approach is analogous to tuning the percentage of the presynaptic neurons that must fire to produce a spike in the post-synaptic neuron of the next layer. A list of percentage firing thresholds for population projects can be found in Ta-

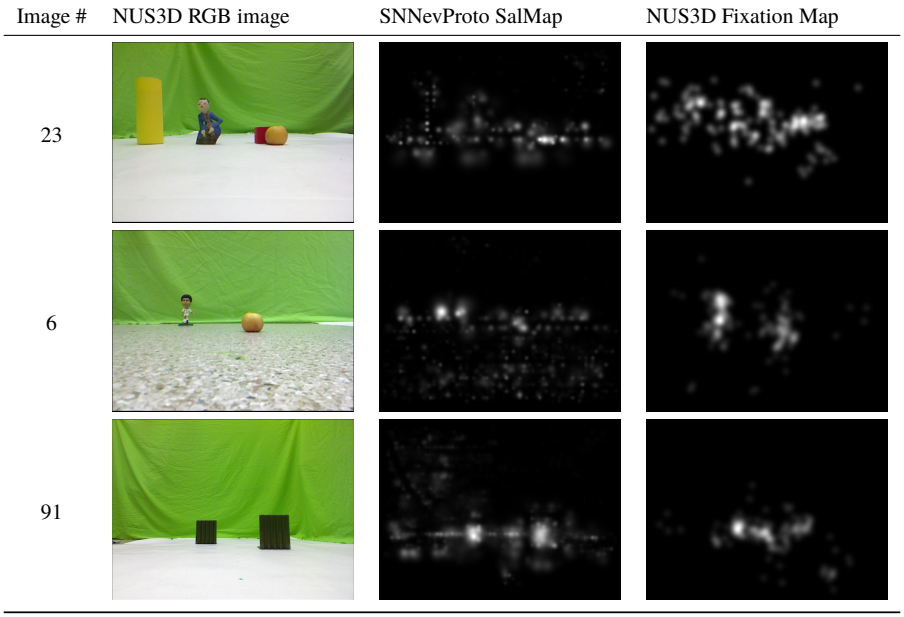


Figure 3.7. Representation of examples from the NUS3D (robot scenario) dataset. The three columns represent the input RGB image, the outcome from the SNNevProto and the related ground truth from the NUS3D dataset. These examples show how the model performs when the observer fixation maps focus on objects. The response from the model is with 60% OL and 0.013 inhibition.

ble 3.2. This stage of the SNNevProto mimics the Border Ownership Pyramid in [9]. A similar process to the Border Ownership in [9] pools the activity of mirrored VM filter orientations into a single neuron. The combined filter neuron has maximal activation at the presentation of a closed surface of the same size as the convolution filter size. Following the Gestalt principles [83] this represents detection of a proto-object. The proto-object spikes are added to a combined saliency map with their energy spread over the surrounding pixels using a 2D Gaussian distribution with standard deviation a third of the filter size in pixels. Therefore, a pooling stage mimicking the Grouping Pyramid is computed making the response size invariant. Values from all scales and the four pairs of rotations are pooled together to produce a combined saliency map.

3.7 Experiments and Results

We validated the SpiNNaker implementation of the proto-object attention model, SNNevProto, by comparing its performance with the PyTorch GPU implementation, PyTevProto.
The system is further benchmarked using the ground truth 2D fixation maps of the NUS-3D dataset [169], obtained recording the eye movements of subjects observing the images of the dataset.

The characterisation compares the responses from the two models qualitatively, show-

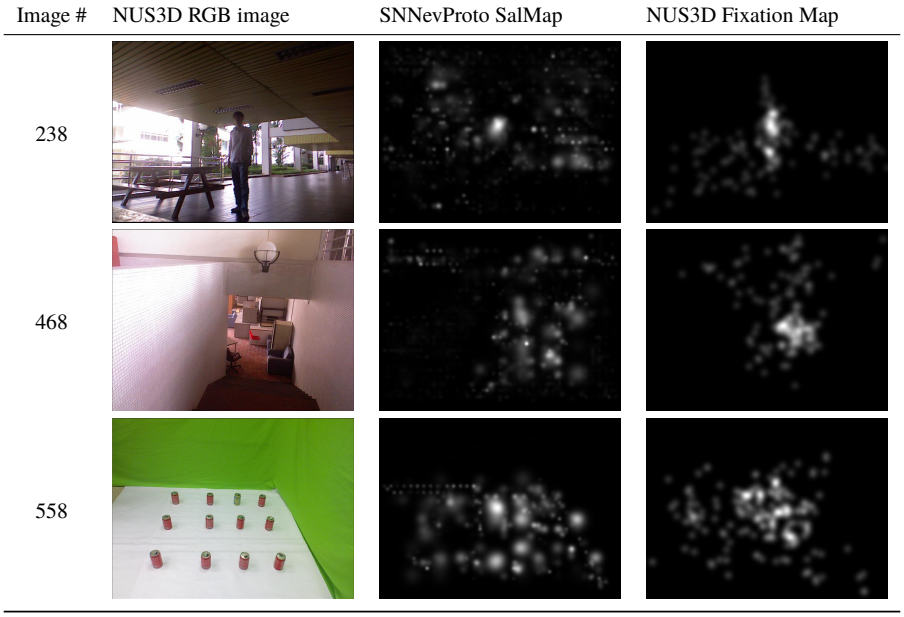


Figure 3.8. Representation of random chosen examples from the NUS3D (random subset) dataset. The three columns represent the input RGB image, the outcome from the SNNevProto and the related ground truth from the NUS3D dataset. These examples show how the model performs when the observer fixation maps are sparse and unclear. The response from the model is with 60% OL and 0.013 inhibition.

dataset #	First sample latency [ms]	Second Sample latency [ms]
1	17	19
2	15	18
3	10	18
4	15	29
5	14	15
6	18	19
7	15	17
8	16	16
9	16	19
10	18	20
11	16	20
12	18	19
13	20	21
average	16$\pm$2.44	19.2$\pm$3.37

Table 3.3. Results of latency in milliseconds for different datasets of SalMapIROS. The test is done measuring the latency of two different samples for each dataset. Each row represents a dataset used to measure the latency in two separate samples. Each dataset represents static and dynamic objects placed in front of iCub (such as a paddle, a puck, calibration circles, proto-object patterns, a mouse, a cup and clutter (see Fig. 3.4)

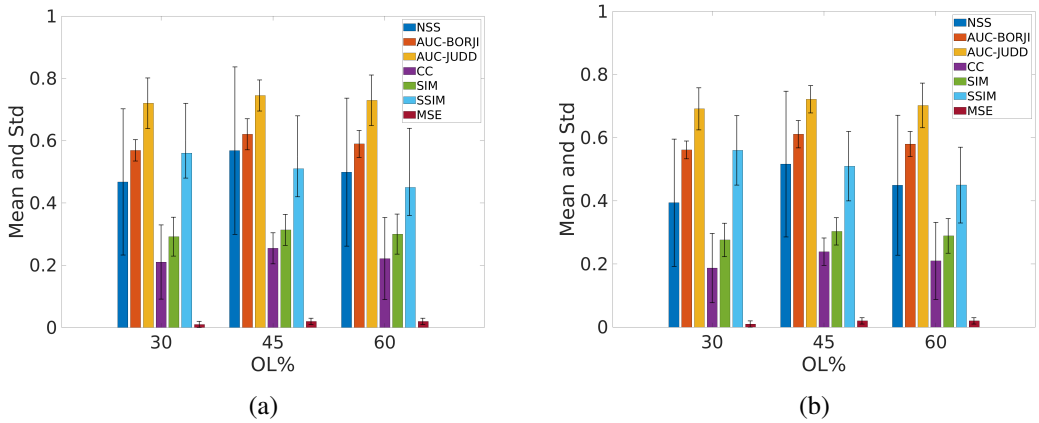


Figure 3.9. Comparison with different metrics evaluating the similarity of the SNNevProto saliency maps with the NUS3D fixation maps (ground truth) [169] in two different subsets (robot scenario (**a**) and random subset (**b**)) for different OL percentages. The metrics used are: the Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji) & (AUC-Judd), Pearson's Correlation Coefficient (CC) and Similarity (SIM) [35], [137]–[139], Structural Similarity (SSIM) and Mean Square Error (MSE). A higher score is better for all excluding the MSE where the lower score determines similarity.

ing the strength and the weaknesses of each system. We then quantitatively compared the response between the SNNevProto and the PyTevProto using the latter model as the baseline. We searched for the best set of parameters, exploring different OL percentages of the VM filters on each layer and the best inhibition value.

To characterise the response, this analysis exploits the SalMapIROS dataset which contains patterns and robotic scenarios with objects and clutter in the scene. The SalMapIROS dataset is obtained recording the events coming from the event-driven cameras mounted on iCub looking at different scenes with real objects or 2D printed patterns. The robot performs small circular periodic stereotyped ocular movements to generate stimulus-dependent activity from event-driven cameras for static scenes. To estimate the selectivity to a range of sizes we used a pattern representing circles of different dimensions (see Fig. 3.4, third row). The other two patterns in Figure 3.4 (first and second row) describe the definition of non proto-object and proto-object exploiting the design used by [9]. The proto-object is represented by the four corners facing each other forming close contours reminding of a square shape. The remaining pictures see objects of different sizes over a desk (fourth row) to study the applicability of our system in a scenario where we want the robot to interact with items in the scene. Figure 3.5 shows two cases of simple clutter represented by a pattern and a bag of nails alongside with an object (a puck).

Figure 3.4 and 3.5 qualitatively show the saliency map from the two models on some samples of the SalMapIROS dataset. Overall, the response from the models is coherent and both implementations detect the objects in the scene. In Figure 3.4 the response from the SNNevProto is less sparse and more localised over the targets which is helpful if a robot

needs to locate and reach the object. The PyTevProto correctly gets rid of the clutter in Fig. 3.5 (first row) but not in Fig. 3.5 (second row). The SNNevProto instead successfully discards clutter in both cases. This results show robustness to clutter of the SNN model. This behaviour was achieved by tuning the level of inhibition. By balancing inhibition appropriately the filter can be made selective to the VM kernel shape without silencing the firing of the filter neurons. As the clutter did not contain the specific contours the VM filter is selective to, the inhibition effectively suppresses firing from the filter neurons.

The SalMapIROS dataset has been used also to obtain data related to the latency measurements. As the SpiNNaker simulation is run in real-time, latency is both walk-clock time and simulated time. The results in Table 3.3 show the amount of time needed to obtain spikes from the proto-object neurons, which compose the saliency map, given an input. Each sample is obtained by waiting for the onset of input spikes following a quiescent period and measuring the time taken for the activity to flow out of the model. This allows the delay of the input spike to the consequential output spike to be most clearly extracted. The average latency is 16 ms (2.44 ms standard deviation) and 19.2 ms (3.37 ms standard deviation)[3] for the second set of samples, compared with the 170 ms needed in average for the PyTevProto model[4] to obtain a saliency map of the scene. Figure 3.6a shows the comparison between the SNNevProto and the PyTevProto saliency maps using the SalMapIROS dataset. We evaluated the similarity among the outcomes using Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji) & (AUC-Judd), Pearson's Correlation Coefficient (CC) and Similarity (SIM) [35], [137]–[139], Structural Similarity (SSIM) and Mean Square Error (MSE). These metrics are computed to compare the saliency maps to the ground-truth, following standard analysis methods in the literature[35], [137]–[139]. A single saliency map cannot perform well in all the metrics since they judge different aspects of the similarity between ground truth and predicted saliency map [140]. These metrics offer a way to determine how well a saliency-based model approximates human eye fixations. The properties of the chosen images for the benchmark, such as dataset bias (centre biasing, blur and scale), probabilistic input and spatial deviations, affect the result of the metrics [137]. Saliency based models can include such properties. In this work the robot needs to detect objects of different sizes to potentially interact with them. In fact, the SNNevProto only focuses on the scale of the objects rather than other properties. MSE and SSIM are metrics used in classical computer vision to explore the similarity among images. MSE estimates the error between two images and it is a global comparison, and the SSIM estimates the similarity between two images taking into account structural changes in the images.

[3]The model is running in simulation on the "big machine" using 6 SpiNNaker boards and 48878 neurons.

[4]The PyTevProto is the PyTorch implementation of the same bioinspired attention model running on a laptop with Nvidia GTX 1650 GPU and Intel Core i7-9750H CPU @ 2.60GHz x 12

There is not a significant difference over the OLs percentages comparing the saliency maps between the SNNevProto and the baseline (PyTevProto). Only AUC-JUDD and SIM slightly increased increasing the OL percentage. Although there is not a remarkable increment we chose 60% OL to explore the inhibition parameter (μS conductances). 60% OL represents a good compromise among the robustness of the model, ensuring enough overlap to cover the whole visual field without losing any area of the visual field, the number of SpiNNaker boards needed (see Table 3.1) and the results obtained. Each significant increment of neurons causes an increment on the number of SpiNNaker boards required. Nevertheless, the number of neurons required does not affect the latency of the model because the pipeline remains unaltered. Figure 3.6b explores a range of different inhibitions showing again not a significant incremental or decremental trend. Only SIM and CC show a slight improvement increasing the inhibition parameter. The results exhibit a stable response exploring different parameters showing no need to create a complex network with a large number of neurons to get usable saliency maps. Overall SSIM and AUC-JUDD seem the best metrics to explain our saliency map results.

Along with the characterisation where we compared the response of our implementation with the PyTevProto, we evaluated the response from the model by benchmarking the saliency maps with the ground truth provided by the NUS-3D dataset [169]. The investigation includes the comparison between the saliency maps generated by the SNNevProto and the fixation maps qualitatively and quantitatively evaluating the similarity between the two maps. The 2D fixation maps of the NUS-3D were collected from subjects looking at images while recording eye movements. The ground truth obtained recording the response from the subjects includes different mechanisms of bottom-up and top-down processings, increasing the complexity of the observers' fixations. The observer response does not exclusively derive from a data-driven process but also a task-driven mechanism driving the gaze towards a particular region of the scene. Attention is a complex interplay between these two mechanisms combining bottom-up and top-down mechanisms to perceive the surrounding [141]. The model we propose is a bottom-up system that does not include top-down mechanisms, but 2D fixation maps can be used to evaluate the response of our system as they represent the only ground truth we can refer to.

To use the NUS-3D dataset within the event-driven proto-object model, we used the Open Event Camera Simulator [176] shaking the images to simulate small periodic circular eye movements.
We chose two subsets of data from the dataset: one is a selection of 50 images representative of a robotic scenario (robot scenario) and the second one is a collection of 50 random images (random subset). The first subset (see Fig. 3.7) represents a simple robotic scenario where objects are placed over a surface. The second subset (see Fig. 3.8) is a random se-

Metrics	
Normalized Scanpath Saliency (NSS)	CC approximation, good for saliency evaluation.
Area under ROC Curve (AUC)	Invariant to monotonic transformations, driven by high-valued predictions. Good for detection applications.
Pearson's Correlation Coefficient (CC)	Linear correlation between the prediction and ground truth distributions. Treats false positives and false negatives symmetrically.
Similarity (SIM)	Similarity computation between histograms, more sensitive to false negatives than false positives.
Structural Similarity (SSIM)	Similarity among images, highly sensitive to structural changes.
Mean Square Error (MSE)	Similarity among images, global comparison.

Table 3.4. Metrics summary. This table takes inspiration from [137]

lection among all the dataset images adding complexity and variety to the scenarios.

Qualitatively, the saliency maps from the model and the fixation maps are sparse and not easily understandable at a first glance (see Fig. 3.7 and Fig. 3.8). Figure 3.7 represents a scenario where the SNNevProto Saliency Map and the ground truth target select the same objects as interesting. The highest response (brightest) is located around the objects in the scene. Figure 3.8 shows a slightly sparse response from the model compared to the fixation maps, not allowing a clear understanding of the agent's attention.

Quantitatively, Figure 3.9 shows good results for both datasets exploring different percentages of OL. Furthermore, all the metrics do not show a significant increment changing the OL%, validating the response of the model either for simple or for complex scenarios.

Although we do not include the complex bottom-up top-down interplay [141] in our implementation, overall the results yield a good representation of the scene for our purposes. Moreover, the metrics used to quantify the similarity do not give equal results among them. All the metrics are used in literature to explain saliency-based model performances. They compare different aspects depending on the ground truth representation and the definition of the saliency map of the model. These metrics treat differently false negative and positives, viewing biases, spatial deviation and the pre-process of the saliency maps. We were initially interested in the location of the responses from the saliency maps rather then the value in that position, choosing the SSIM as the metric we could rely on. SSIM

estimates the structural similarity between two images comparing small sub-samples of the images with each other. This metric well describes our situation where we are more interested in having a response in the same location rather then having the same amount of response in terms of intensity. We further added other metrics used in literature for completeness [137]. The results seem to bare out our expectations. Overall, in our case SSIM seems a good metric to explain our saliency maps. Alongside with the SSIM, AUC-JUDD provides good results too, where each saliency pixel is treated as a classifier splitting them in ”fixation” and ”background”. This metric computes the ratio of true and false positives to the total number of fixations and saliency map pixels using a thresholded mechanism [177].

3.8 Conclusion

Overall the response of the spiking implementation of the event-driven attention model on SpiNNaker (SNNevProto) is coherent with the PyTevProto, showing a significant improvement in removing the clutter with respect to the baseline GPU-based implementation (PyTevProto). This can be well explained by the nature of the model. The SNN model, as a result of the inhibitory connections, is far more selective to the shape of the VM filter, than in a classic convolution using a kernel with no negative weighting. The convolution will produce activity everywhere the filter overlaps with events, enabling clutter to evoke a response in the saliency map. The advantage of the resulting higher selectivity and localised activity in the saliency map is in the possibility to improve object localisation and segmentation and, hence, the interaction of the robot with the selected object.

For the same structural reason, the response from the SNN is less sparse and focused on the location where the detected objects are placed. Two VM filter of opposite side are connected together at every scale and with different rotations. Only when they both respond there is a response from the successive layer of the SNN. Therefore, this significantly helps in generating a preciser saliency map.

Given the parallel structure of SpiNNaker, increasing the number of neurons does not affect the latency performance. For this reason, we tested the model for increasing the OL percentage, and therefore increasing the density of the convolutions. This strategy appears to provide little benefit to model performance and requires the use of an additional number of SpiNNaker boards. Results for low values of OL percentage, equivalent to a large stride in CNNs, produce a similarly reasonable representation of the visual scene compared to high values, with significantly reduced network size. This displays the feasibility of fitting

the SNNevProto model on a single SpiNNaker board and having it work in tandem with the iCub humanoid robot.

The SNN implementation provides a saliency map of the scene in around 17.5ms. In comparison with the PyTevProto (120ms), these results are a significant improvement, that enables the system to run online in dynamic environments, where the saliency map can be used to drive the gaze and actions of the robot in real-time. To this aim, the SNN implementation on SpiNNaker could easily include Winner-Take-All competition and Inhibition of Return [75] to dynamically select the location of the next saccade of the robot. Additionally, the saliency map allows the system to focus its attention towards a specific target, devoting computational resources to perform other tasks, such as object recognition, only in the area where they are needed.

Finally, the attention and gaze of robots are extremely important in the interaction with humans [178], we, therefore, questioned how close the saliency map (used as a proxy for the robot's fixational eye movements) was close to humans. We validated and characterised the system, but the quantitative results of the benchmark do not capture the true merit of the model. Quantitatively, the similarity among the benchmark results (robot scenario and random subset datasets) suggests another question; how do we define the complexity of a scenario? and which aspects should we take into consideration for attention? These results proved to us that the random subset does not produce lower results, hence, it may not contain as complex scenarios as we expected. Each metric captures a specific aspect of the saliency maps, our analysis is instrumental to give a quantitative comparison but mostly to study the effects of the different parameters on the model performance. Moreover, most of the metrics present a high variance due to the mismatch between the SNNevProto saliency maps and the ground truth. This should be investigated in depth creating several subsets from the 600 images of the NUS3D dataset investigating the variability of the response. As expected, a pure bottom-up neuromorphic attention system taking into consideration only the intensity as a feature to determine the saliency map only partially predicts the fixational eye movements of humans. To this aim, the model can be enriched with other channels (such as motion, depth, texture, etc) and with top-down processing to focus the attention towards a specific task.

The model could benefit from the leveraging of learning dynamics in the fine-tuning of network parameters. This could allow the model to adapt itself to particular data sets and reach a higher level of performance. This may improve the inference of the model given appropriate training and data as compared to handcrafted parameter selection.

Moreover, the spatial integration [179] and the lateral inhibition [180] could enrich the

model following a detailed bioinspired pipeline and further reduce the amount of data to be processed. Finally, further experiments could be done emphasising the clutter removal capabilities and exploring the potentiality of the model.

3.9 Acknowledgements

The development of SpiNNaker was supported by EPSRC grants EP/D07908X/1 and EP/G015740/1 and ongoing software development is supported by the EU Human Brain Project (H2020 945539). This project has been possible thanks to the HiPEAC European network grants 2019. We also would like to show our gratitude to Jay Perrett for sharing his accurate review and support during the implementation of this work.

3.10 Reflections & Conclusions

The SNN version of the saliency-based proto-object model clearly detects proto-objects in different environments especially removing clutter as a possible interesting feature. Further experiments investigating the clutter removal could be interesting for a further more complex application along with the figure-ground segmentation concept [181]
This pipeline is fully neuromorphic and gets rid of the event-frames representation taking advantage of the spike-based architecture. The impressive result of this implementation is due to the use of the neuromorphic platform, SpiNNaker, reducing dramatically the overall system latency. Given the first spike from the event camera, the first spike from the saliency map is produced after only 4 ms. This latency is computed as the first spike-in/spike-out of the system. Therefore the first output spike could not represent the most salient object in the scene. A more accurate investigation should address the real system latency since the stimulus presentation (onset) to the first stable saliency map.
Smart connectivity in the network allowed the replacement of several layers of processing and ensures an asynchronous response.
This work includes the investigation on the quantitative results validating the model through comparison methods used in literature. The saliency map computed from the network is sparse due to the spike-based implementation, still achieving a reasonable comparison with the ground truth. The results changing the overlap among filters did not show any substantial change in the response, suggesting the choice of simple scenarios as a used dataset (SalMapIROS). The intuition would see a change in response among overlap percentages exploiting a more crowded scene requiring a high number of VM filters on the input layer. These results do not see a 0% overlapping percentage because of the bio-inspired choice

to ensure robustness [173], [174] for the model. The inhibition results show the same, not informative conclusions, confirming the possible not adequate dataset choice for these investigations. The explored inhibition values represent the empirical range of values found to generate a clear response from the system. The quantitative investigation on the benchmark dataset allowed a better understanding of the results providing a new set of questions around the choice of the input scenes. The saliency maps obtained from the ”robot scenario” dataset show the same quantitative results as the ”random subset” dataset suggesting a wrong splitting of the two datasets erroneously considering the ”random subset” as the complex scenario.

Although the quantitative evaluation provides a quantifiable measurement of closeness to the ground truth, the internal cognitive biases from subjects looking freely at an image do not guarantee an absolute comparison with the saliency maps. This problem could be solved by defining a clear task (i.e. focusing on the closest target as further done in Chapter 4). The complex interplay between bottom-up and top-down human mechanisms does not allow for a fair comparison with the purely bottom-up outcome of the proposed model. The two outcomes cannot indeed be identical per se. Moreover, the metrics look at different aspects of the similarity, making the perfect match with the ground truth further impossible. AUC-JUDD and SSIM are described to be the metrics to explain the results obtained. Both metrics are driven by the high response in the same location, being sensitive to structural changes. These metrics do not explain the results, but they are only the best scores obtained among the used metrics. To the current author's knowledge, these experiments are non-informative, suggesting a more detailed investigation to characterise the model response for each metric using ad-hoc datasets.

Despite the clear advantages of this neuromorphic implementation, the number of required SpiNNaker boards increases significantly with the percentage of filter overlap. Maintaining a good percentage of overlap (60%) the system needs up to 48878 neurons and 6 boards in parallel running remotely on the ”big machine” exploiting batch processing.

However, the quantitative results obtained have shown a good response to the model reducing the overlap percentage from 60% to 10%. This percentage of overlap still requires 3 SpiNNaker boards and 10428 neurons. A possible online implementation on iCub, using only one board for convenience, would force a simplification of the pipeline and a considerable reduction of the visual field. This consideration confirms the scalability problem on neuromorphic platforms mentioned in the introduction.

Whether the platform would allow an increased number of neurons on the physical board this SNN implementation of the saliency-based proto-object model could certainly replace the PyTorch implementation.

Chapter 4

Event-driven Proto-object based saliency in 3D space to attract a robot's attention

The supplementary materials can be found in Appendix A, section A.

This work has been published.

Ghosh, S., D'Angelo, G., Glover, A., Iacono, M., Niebur, E., & Bartolozzi, C. (2022). Event-driven proto-object based saliency in 3D space to attract a robot's attention. Scientific reports, 12(1), 1-14.

4.1 Personal Contribution

Theoretical Implementation	Code development	Experiments design	Experiments execution	Analysis of the results	Paper writing	Academic authorship
yes	yes	yes	yes	yes	yes	First co-author

4.2 Authors

Suman Ghosh, Corresponding First Co-Author, Istituto Italiano di Tecnologia
Giulia D'Angelo, Corresponding First Co-Author, Istituto Italiano di Tecnologia - The University of Manchester
Arren Glover, Istituto Italiano di Tecnologia

Massimiliano Iacono, Corresponding First Author, Istituto Italiano di Tecnologia
Ernst Niebur, Johns Hopkins University
Chiara Bartolozzi, Istituto Italiano di Tecnologia

4.3 Authors Contribution

C.B. conceived the main idea behind the work. S.G. and G.D. developed the theory for the disparity estimation and its integration with the event-driven proto-object model. S.G. developed the entire pipeline on the robot, with help from A.G. M.I. implemented the event-driven proto-object model in PyTorch used in this work. S.G. and G.D. designed the experiments with supervision from C.B.; S.G. and G.D. conducted experiments. S.G., G.D. and C.B. analysed the experimental results. S.G., G.D. and C.B. wrote the manuscript. A.G. and E.N. gave valuable feedback and helped edit the manuscript. G.D. made the supplementary video accompanying this work.

4.4 Abstract

To interact with its environment, a robot working in 3D space needs to organise its visual input in terms of objects or their perceptual precursors, proto-objects. Among other visual cues, depth is a submodality used to direct attention to visual features and objects. Current depth-based proto-object attention models have been implemented for standard RGB-D cameras that produce synchronous frames. In contrast, event cameras are neuromorphic sensors that loosely mimic the function of the human retina by asynchronously encoding per-pixel brightness changes at very high temporal resolution, thereby providing advantages like high dynamic range, efficiency (thanks to their high degree of signal compression), and low latency. We propose a bio-inspired bottom-up attention model that exploits event-driven sensing to generate depth-based saliency maps that allow a robot to interact with complex visual input. We use event-cameras mounted in the eyes of the iCub humanoid robot to directly extract edge, disparity and motion information. Real-world experiments demonstrate that our system robustly selects salient objects near the robot in the presence of clutter and dynamic scene changes, for the benefit of downstream applications like object segmentation, tracking and robot interaction with external objects.

Multimedia Material

`Video:https://zenodo.org/record/5091539`

4.5 Introduction

Every agent, whether animal or robotic, needs to process its sensory input in an efficient way, to allow understanding of, and interaction with,the environment. Since the agent's computational capabilities are limited, careful allocation of perceptual and cognitive resources is required [182]. The process of filtering relevant information out of the continuous bombardment of complex sensory data is called selective attention. This process not only occurs in animals, where the selection of the most ecologically important stimuli like the presence of a predator is required but also in complex machinery with a rich array of sensors, like robots. The large amount of information arriving in the information processing stages at **all** times from sensors that are needed only at **some** times cannot be processed economically in its entirety. Selective attention mechanisms are used to analyse only the most important subset of the sensory stream. A number of visual attention algorithms have been proposed in robotics exploiting selective attention mechanisms [44], [46]–[49].

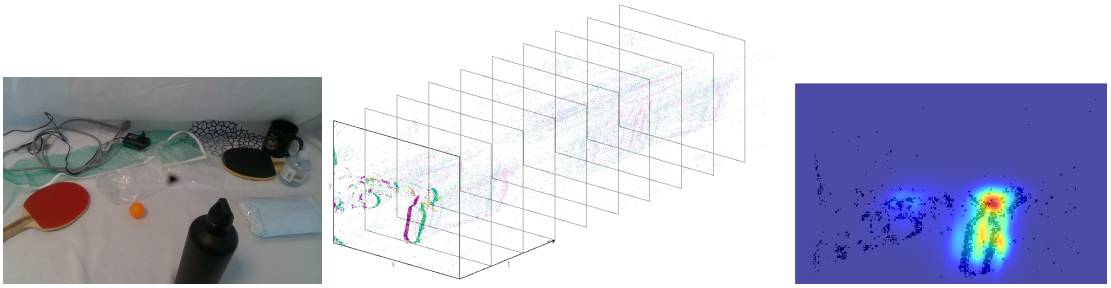


Figure 4.1. Event-driven proto-object saliency estimation in 3D. Left: Cluttered table top with objects of different sizes and textures placed at varying depths (only for visualisation). Middle: Events produced from the ATIS camera using circular robot eye motion. The event stream is plotted in spatio-temporal coordinates. The green and purple colours represent whether the pixel witnessed a brightness increase or decrease. Events from the stereo cameras serve as input to our model. Right: Saliency map computed using the proposed evProtoDepth model, with the closest object (black bottle) selected. The saliency map is overlaid on the event image generated by accumulating events generated within a 100 ms time window.

Visual attention is the result of the complex interplay between the physical characteristics of the scene (stimulus-driven, bottom-up mechanisms) and the goals of the agent (task-dependent, top-down mechanisms) [183]. Bottom-up models of selective attention rely both on feature extraction [102], [147], [184] and perceptual organisation of the scene [9]. Mechanisms of perceptual organisation have been formalised in the form of "Gestalt laws"

(*e.g.* continuity, proximity, figure-ground segmentation) that contribute to the grouping of visual features into coherent objects [83]. These principles can be integrated into feature based bottom-up models [33], [75] to identify so-called proto-objects [9], by adding a layer of Gabor [102], or curved Von Mises filters [9], loosely similar to neuronal responses in primate visual cortex [98]. Such models use biological inspiration by emulating the cells that extract visual features and combine them using border ownership and grouping mechanisms, to produce a robust saliency map of the scene that increases perceptual saliency of regions with object-like stimuli.

We are interested in the bridge between biologically plausible models, bio-inspired hardware, and embodied agents (robots) to further understand the role of the hardware and the environment in selective attention processes. Our previous work [166] implemented the proto-object model proposed by Russell et.al. [9] using bio-inspired artificial visual sensors, called event-driven cameras [5]. The event-driven cameras function more similarly to biological eyes than frame-based cameras. Instead of scanning each pixel in order to measure the incident light level as in a traditional camera, each pixel in an event camera is independent and produces a spike when the incident light changes beyond a threshold. These ”pixel spikes” are similar in function to the action potentials that the retina sends to the brain. The output of the event-camera is asynchronous, sparse, and occurs only where there is a differential between dark and light regions of the scene detected as an illumination change of each pixel over time, functioning *de facto* as a dynamic edge extractor. The integration of the event-camera into the proto-object processing pipeline inherently performs some of the lower-level processing that the model requires (detecting illumination change), opening interesting questions on the role of the hardware, as well as the brain, in sensory processing.

Relative depth and apparent object size provide important cues to guide bottom-up attention mechanisms during physical scene interpretation [185]–[187]. Depth cues from binocular disparity have been shown to modify eye movements of participants when shown 3D images [188] and videos [189]. Directed attention to local features have also been shown to aid in the interpretation of three-dimensional cues [45]. To explore the role of depth in event-driven attention, in this paper, we extend our previously developed event-driven proto-object model (evProto) [166] by combining it with a biologically inspired stereo disparity estimation algorithm [11], resulting in a depth-based attention model. Furthermore, our implementation runs online on a robotic platform (the neuromorphic iCub [2]).

In two previous studies, a proto-object based model of selective attention[9] was extended to include depth in the saliency map computation [88], [109]. Our model goes be-

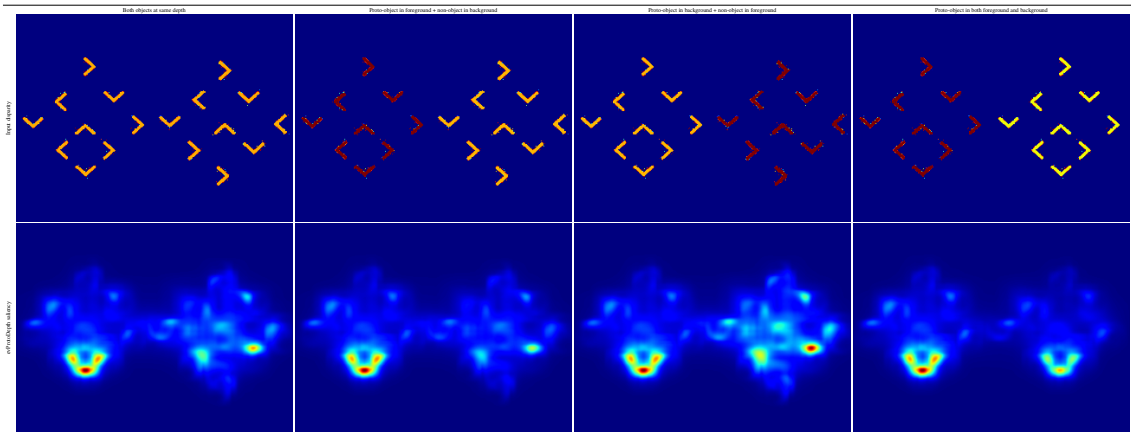


Figure 4.2. Interplay between depth (disparity) and Gestalt cues in evProtoDepth saliency. The disparity maps (Row 1) have two possible depths: near (dark red) and far (light orange), and the evProtoDepth saliency (Row 2) is shown from strong (red) to weak (blue). Arranging the angle features in a convex shape generates a perceptual (proto-)object that contributes to saliency in our model. Turning any of the angles in a different orientation destroys object perception. This contribution to saliency is integrated with that resulting from differences in depth. The salience of the synthetic proto-object pattern increases as it moves closer to the camera. However, even when the proto-object moves further away in the background, it produces a strong response compared to the non-object pattern in the foreground. This demonstrates the advantage of using a proto-object model instead of directly relying on raw scene depth for nearest "object" selection by the robot. The selectivity is the strongest when the proto-object is placed closer to the camera while the non-object pattern is in the background (Column 2).

yond those studies mainly in two ways. First, both of these models are frame-based while we use input from neuromorphic event cameras. Second, both models require supplementary information in addition to the two input images. A full depth map obtained by an RGB-D sensor is needed for the Hu et. al.odel [88]. The Mancinelli et. al.odel [109] does obtain depth information from stereoscopic cameras but it assumes that a certain number of known correspondence points are available. Instead, our model solves the correspondence problem directly, using only visual input streams from two event-driven cameras by making use of the precise signal timing at the pixel level, as is described in Methods.).

An important concept for all agents interacting with their physical environment, be it humans, animals or robots, is the implicit, underlying interpretation of the environment imposed by object affordance [190]; the object features that define their possible uses and/or make clear how they can or should be used [191]. It seems reasonable to expect that there is a bi-directional relationship between affordances and salience: Affordances are important for interacting with objects, so they need to be attended to make this interaction possible. On the other hand, features related to affordances may be salient by themselves, either by their inherent visual properties (shape *etc.*) or by their design (*e.g.* painting a handle red). There is evidence for a bidirectional relationship between attention and affordance [192], [193], while other studies have shown that the correlation may not be particularly strong [194]–[196]. The relationship may be more nuanced and be affected by

additional neurological systems, which would require additional study. We note, however, that even though we do not include any explicit consideration of affordances in our study, we direct the robot's attention towards objects in a certain size range, according to its grasping capabilities [166]. Furthermore, in our implementation of the depth channel we increase the saliency of closer objects, which are therefore easier to reach by the robot, which is an affordance of elementary importance.

Our motivation is to understand the benefits of combining biologically inspired algorithms with neuromorphic hardware on embodied agents, as opposed to improving the precision and performance of the object selection or eye fixation prediction. The objective we pursue with our attention model is to produce saliency maps that are robust to noise, quickly adapt to dynamic changes in the visual scene, and remain close to important biological processing mechanisms. As the system produces saliency estimation using event-driven cameras based on depth information, we will refer to it as the evProtoDepth (event-driven Proto-object 3D) model. It is able to cope both with dynamic scenes (with motion) and with static images. In order to process the latter, small periodic stereotyped ocular movements are performed by the robot to generate stimulus dependent activity from event-driven cameras to generate pixel motion, akin to microsaccades in biological vision[158].

Since we want the robot to be more attentive to nearby objects that are within its reach, our saliency model design puts a higher importance on stimuli with higher disparity. This allows nearby objects to inherently appear more salient. Besides the affordance of reachability, our design choice is also based on ecological evidence which suggests that attention in insects, mice and humans is drawn towards looming stimuli [197]–[199], wherein nearby approaching objects are deemed especially important. Whereas other features also contribute to salience in full attention systems, here we focus on depth alone and leave the integration with other submodalities for future work. Thus, the evProtoDepth model selects the nearest potential object (proto-object) that the robot could reach and interact with as the most important item in the scene (see Fig 4.1). To fully explore the influence of depth on the event-driven saliency model, we propose a depth-only implementation as the base for a more complex saliency based attention system in the future, in which multiple features are weighted based on top-down mechanisms to adapt the detection of salient regions of the scene to the task at hand [200].

In the next section, we demonstrate the performance and suitability of the event-driven stereo depth algorithm as an input to the proposed attention model. A comparison of the proposed evProtoDepth and the non-event-based proto-object attention model is made on publicly available attention-based datasets, and a series of tests on the iCub robot are made

to demonstrate the attention to nearby objects, as opposed to nearby non-objects and far away objects.

	NSS mean, median	AUC_Borji mean, median	KLDiv mean, median	CC mean, median	SIM mean, median
fbProtoDepth [88]	**0.936**, **0.917**	**0.737**, **0.747**	**1.603**, **1.501**	0.386, 0.386	0.283, 0.296
evProtoDepth	0.769, 0.888	0.606, 0.592	2.971, 2.249	**0.417**, **0.417**	**0.401**, **0.420**

Table 4.1. Consolidated MIT saliency metrics Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji), Kullback-Leibler Divergence (KLDiv), Pearson's Correlation Coefficent (CC) and Similarity (SIM) [35], [137]–[139] on the closest-object subset of the NUS3D dataset. A higher score is better for all metrics, excluding the KLDiv. Bold font indicates the model with the better performance. Some of the corresponding scenes and saliency maps are depicted in Fig 4.3. The metrics for each individual image in this subset are presented in Supplementary Fig A.8.

4.6 Results

The evProtoDepth model is biased to select the closest object in the scene, and to decrease the saliency of near stimuli that do not fulfil the continuity and proximity conditions that define the presence of a proto-object, as shown in Fig 4.2. We evaluate the evProtoDepth model against the standard frame-based proto-object model (fbProtoDepth) [88] on a subset of the NUS3D publicly available dataset [169], comparing also to ground truth fixation maps captured from human eye tracking data. We validate the accuracy of the online event-based depth estimation model and on the neuromorphic iCub robot [2] with live visual data from stereo ATIS cameras [5], and evaluate the response of the full evProtoDepth pipeline on the iCub robot to identify salient regions produced by nearby objects in the scene.

The model takes $\approx$ 170ms to compute saliency of one frame on a laptop with Nvidia GTX 1650 GPU and Intel Core i7-9750H CPU @ 2.60GHz x 12[1]. The parameters used to run the model are specified in the supplementary material (Tables A.2 and A.1)[2].

An accompanying video (`https://zenodo.org/record/5091539`) supports an intuitive understanding of the experiments.

4.6.1 Saliency Benchmarking with NUS3D Saliency Dataset

The NUS3D dataset [169] is used to quantitatively compare event-based evProtoDepth with frame-based fbProtoDepth against a ground-truth saliency map. The goal of the

[1]The disparity extractor is implemented on C++ and the visual attention proto-object model is a PyTorch implementation
[2]Supplementary materials can be found in Appendix A

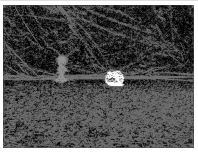
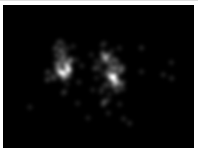
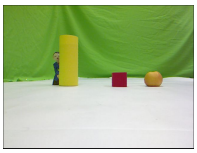
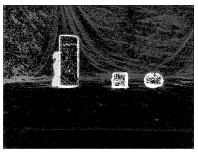
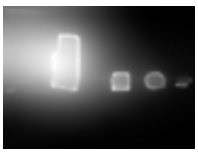
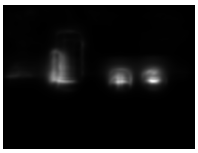
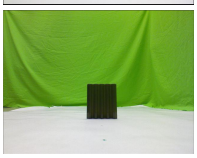
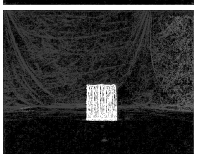
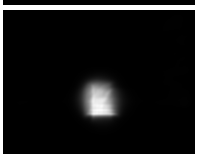
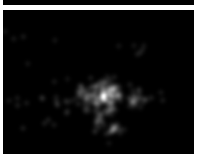
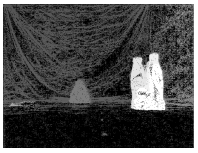
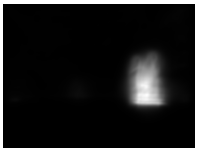
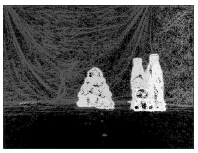
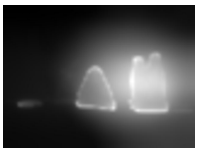
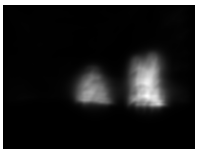
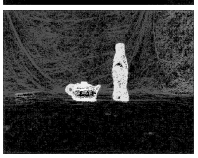
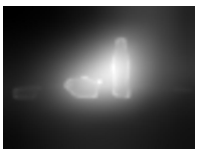
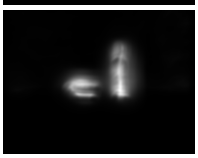
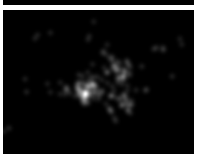
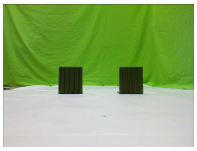
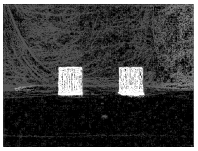
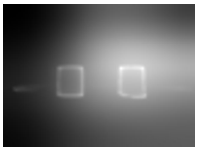
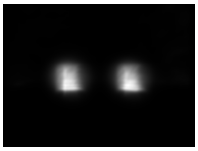
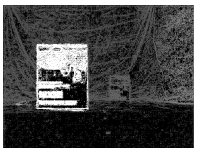
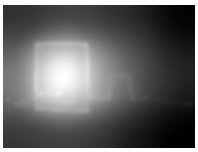
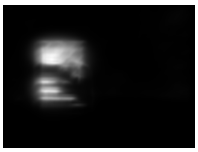
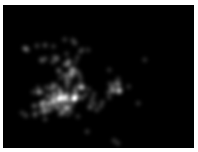
Image #	RGB image	Simulated event disparity maps used as inputs to evProtoDepth	Saliency Map: fbProtoDepth [88]	Saliency Map: evProtoDepth	Ground truth 3D fixations
6					
20					
54					
78					
82					
87					
89					
101					

Figure 4.3. Comparison of saliency maps generated by fbProtoDepth [88] and evProtoDepth on samples from a subset of the NUS3D dataset where ground truth fixation was concentrated on the nearest object in the scene. The subset comprises all cases where the cross-correlation between the ground truth 3D fixation and inverse of ground truth depth ≥ 0.5. These scenes depict scenarios relevant to a robot application where the goal is to select the nearest "object". This benchmarking experiment investigates how depth contributes to event-based proto-object saliency for predicting human eye fixations in such scenarios. evProtoDepth uses a single depth channel for saliency prediction whereas fbProtoDepth combines information from parallel depth, colour opponency, intensity and orientation channels at the final stage. This causes the former to generate sparser saliency maps highly localised on the nearest object which are suitable for robot applications like segmentation and grasping.

analysis is to understand how close we are to the real fixation maps in cases where humans fixate mostly on the closest object. To this aim, we algorithmically selected a subset of **19** images from the dataset in which the highest salient region should be the closest object, *i.e.* images in which the cross-correlation between the ground truth fixation map and the inverse of ground truth depth is ≥ 0.5.

The dataset provides colour RGB input stimuli, depth maps as well as locations of fixations when humans fixated on either the 2D or 3D images. To produce simulated "micro-saccades" (see above), the still images were shifted by 1 pixel in the cardinal directions (right, left, top and bottom) to simulate random small eye motion [201] and a video of 50 frames (25 fps) was created for each input image. Events were generated from the video using the Open Event Camera Simulator [176]. Depth was assigned to each event using the ground-truth depth map for each pixel and smoothed by 1 pixel in each direction to account for the eye-motion. The evProtoDepth saliency map is computed from the simulated events whereas the fbProtoDepth is computed from the static RGB and depth images in the dataset. Fig 4.3 shows that both models detect the objects in the scene focusing the attention on the closest one. The fbProtoDepth shows a wider and centre-biased response, whereas the evProtoDepth shows a more localised response which is useful in a robotic context. It allows the robot to pinpoint the location of most salient parts of the scene with higher precision and confidence, which is important for subsequent physical interaction.

The Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji), Kullback-Leibler Divergence (KLDiv), Pearson's Correlation Coefficient (CC) and Similarity (SIM) are computed as metrics to compare the saliency maps to the ground-truth, following standard analysis methods in the literature[35], [137]–[139]. A single saliency map cannot perform well in all the metrics since they judge different aspects of the similarity between ground truth and predicted saliency map [140].

The fbProtoDepth model has better performance than evProtoDepth on three of the five metrics (NSS, AUC-Borji, and KLDiv), while evProtoDepth achieves a better result for the CC and SIM metrics as shown in Table 4.1. The fbProtoDepth model uses intensity, colour, and opponency channels, while evProtoDepth uses only the depth channel, and as such saliency patterns are not expected to be identical between methods.

The response of both models and the ground truth all peak on the closest objects, as shown in Fig 4.3. While there is not a large amount of clutter in the dataset, it is clear from the second column of Fig 4.3 that the intensity gradient of the background (curtain) is non-negligible and produces many background events. The signal from the background does not conform to the proto-object pattern and therefore is correctly suppressed by the models. In cases in which there is a large difference between the object depth, all models

successfully produce a stronger response to the closest object.

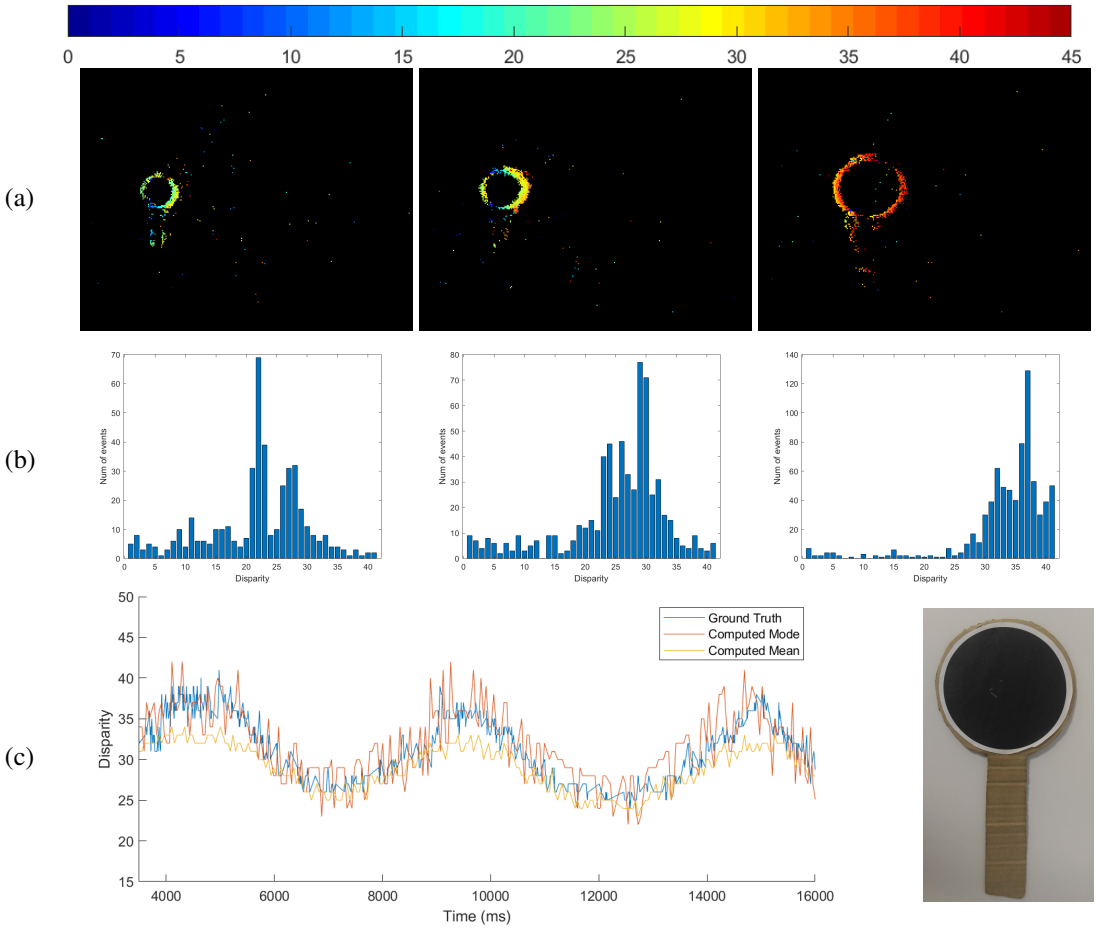


Figure 4.4. Evaluation of estimated disparity accuracy of a circular paddle moving to and fro along the depth axis. a Colour-coded (red=near, blue=far) event disparity maps (time window 100ms) of the paddle at three time instances: far (left), intermediate (centre) and close (right). b The corresponding disparity distribution histograms. c Variation of ground truth and computed disparity (mean and mode within manually annotated ROI) over time, and an image of the input stimulus.

Even in scenes where the ground truth 3D eye fixations were not necessarily confined to the nearest "proto-object", the event-driven evProtoDepth model may produce saliency maps concentrated on the nearest object following Gestalt principles, because it relies on depth information. By contrast, fbProtoDepth, which relies on multiple information channels besides depth, better predicts eye fixations. Some examples of such scenes are shown in supplementary Fig A.9[2].

4.6.2 Disparity Estimation for the Neuromorphic iCub

The accuracy of the disparity estimation model is demonstrated online (50 microseconds latency per event) on the robot by moving a high-contrast fiducial marker, a circle shape, at different distances (within a 30 to 210 cm range) from the stereo cameras and comparing

the computed disparity to the ground truth. The ground truth is computed by tracking the circle shape [202] independently in each camera and computing the horizontal distance between the circle centres in the left and right cameras.

The ground truth is compared to the mean and mode of estimated disparity values within a Region of Interest (ROI) placed around the tracked circle centre. Fig 4.4 shows accuracy of disparity estimation qualitatively and quantitatively. The histogram peaks position in Fig 4.4b corresponds to the depth of the stimulus shown in 4.4a. Fig 4.4c shows quantitatively that the estimated disparity is accurate with respect to the ground truth throughout the sequence. The jitter is due to imperfect time correspondence in the asynchronous system.

Further experiments with more complex multi-object stimuli are presented in the supplementary material (Fig A.7)[2]. We observe that even the noisy disparity map manages to reflect the real scene depth to accurately represent the dynamic environment. The network simultaneously encodes different levels of disparity information, solving the correspondence problem, at different spatial locations and times, consistent with real-world depth. The model is capable of resolving the depth of complex stimuli like the human body, with multiple non-rigid moving parts.

4.6.3 Robot application of 3D proto object model

To validate the evProtoDepth model, we implemented a robot application where iCub uses its movable stereo event-driven cameras to observe static and moving stimuli and selects the nearest proto-object with the goal of further physical interaction. Specifically, we tested whether the evProtoDepth implementation consistently selects the nearest object in the scene as the most salient, when the depth of objects changes dynamically. At the same time, an important aspect of our evaluation is the stability of object selection when the scene configuration remains constant, and the model's robustness to noise both in the background and foreground.

Fig 4.5 shows how the addition of depth information improves object selection stability. The 2D histogram of saliency maps (bottom two rows) obtained during each object configuration shows that both models can select plausible objects in the scene. The addition of the disparity information in the evProtoDepth model, however, enhances the salience of the object which is closer to the observer. While the 2D evProto model assigns overall similar saliency values to the bottle and the mug. The development of saliency over time is shown for both models in Fig 4.6. The comparison between Fig 6d and its evProto counterpart (Fig 6e) shows that the peak response of the evProto model jumps from one object

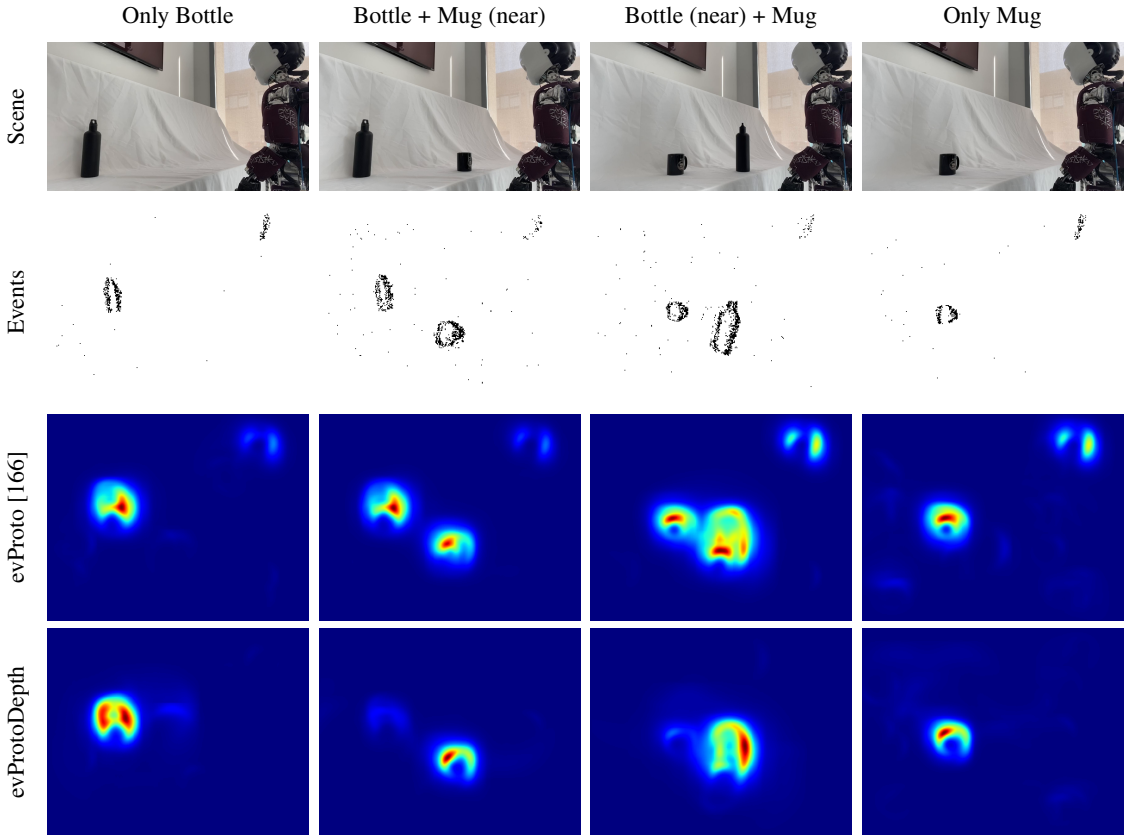


Figure 4.5. Static Objects at Changing Depth *(bottle-mug)* dataset. The columns show snapshots from the 4 different configurations in the dataset. Row 1 depicts the scene from a third perspective. Row 2 shows the event images accumulated over 100ms. Row 3 and 4 illustrate the output generated by the evProto and the evProtoDepth models respectively. Each plot shows the 2D histogram of saliency maps accumulated over all frames in a single input configuration. Object selection is more stable with the 3D model in the presence of multiple objects.

to the other even when the scene configuration remains unchanged. Furthermore, the peak of the saliency map obtained from the evProto model often occurs outside the annotated object boundaries (green dots, ”No selection”). As an example, in Fig 4.5 Row 3 shows that the evProto model finds the ray of the sunlight on the top-right corner of the wall (as seen in the colour image in Row 1) as highly salient. Object disparity therefore stabilises object selection (see Fig 4.6b). The selection does not depend on the number of events generated by the object, as plotted in Fig 4.6c: during the ”Bottle + Mug (near)” configuration, the evProtoDepth selects the mug which is closer to the camera, even though both objects generate similar number of events.

The experiment of Fig 4.7 investigates the response of the system to continuously changing stimuli, in the example shown, a person alternately moving the left and right hand towards and away from the iCub. The location of attention quickly and reliably shifts to the nearest proto-object as soon as the relative position of the hands change sign. The rightmost column shows the location of maximum salience over time, confirming the

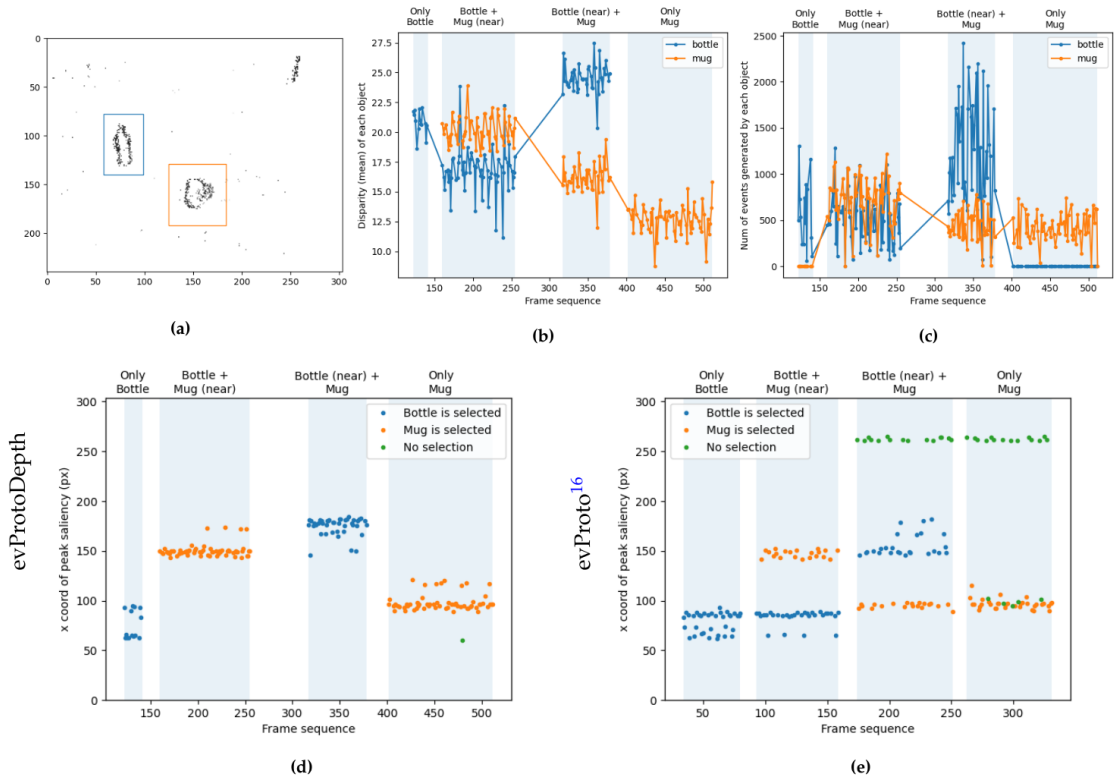


Figure 4.6. The disparity channel stabilises object selection in the bottle-mug dataset, during which the object positions are moved on a table. (a) Sample event frame with manually annotated object boundaries – at this particular time, the mug is closer to the camera. (b) Mean disparity within each object boundary in both object frames. (c) the number of events generated within each object boundary in both event frames. (d), (e) x coordinates of the peak response in each frame for the evProtoDepth and evProto attention models. For each frame, an object is "selected" if the peak saliency pixel lies within its annotated boundaries, otherwise "No selection" occurs. There is only one unique object selection at each time stamp (frame). This means that for evProto in (e), the saliency peak jumps from one object to another frequently. Thus the orange, blue and green dots occur at (different) timestamps very close to each other.

switch of attention from the left hand to the right one while the hands were moving, even when the eyes of the robot are moving. In this second scenario, events are generated by the moving cameras from static objects, leading to high saliency at intermediate depth locations as well (*e.g.* the face of the person standing in front of the camera). However, most of the time the closest objects are selected. This experiment demonstrates that the evProtoDepth model can in real-time track the closest object in a dynamic scene with eyes fixed and in motion.

To obtain a fair comparison between our implementation and the fbProtoDepth model, we recorded RGB-D frames from a Real-Sense D435 depth camera that uses active IR stereo technology to record depth information along with visual images. The depth maps were post-processed with hole-filling filters provided in the Real-Sense library. These holes are 0 value pixels which would otherwise be erroneously treated as the nearest stimuli

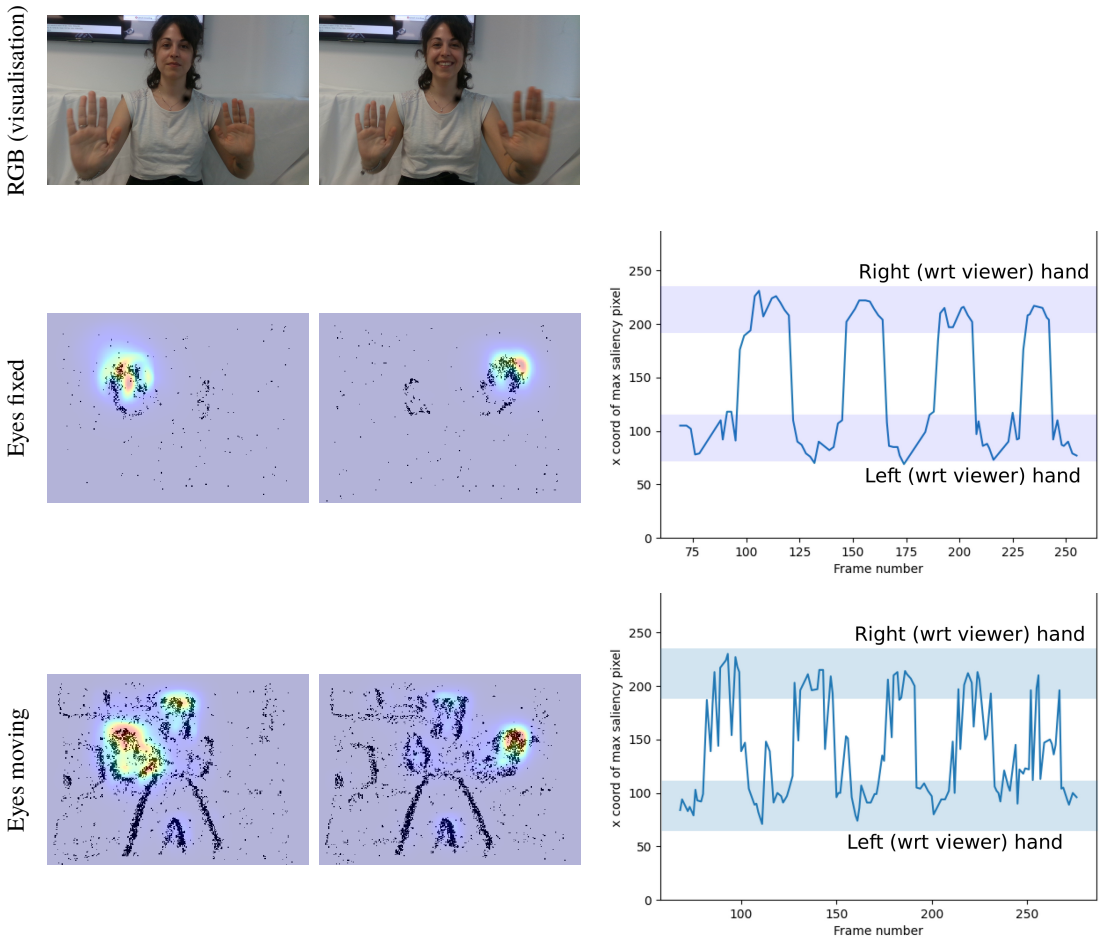


Figure 4.7. Saliency prediction from evProtoDepth in a dynamic scene (data set *hands*) containing hands moving towards the cameras and away from them, with and without the iCub eye motion. The events from the stereo cameras are the only input to our model. RGB (Row 1) frames at different instances of the sequence are shown for visualisation. The two leftmost columns of Rows 2 and 3 depict corresponding saliency outputs overlaid on input events while the robot eyes were fixed (Row 2), and moving (Row 3). With *fixed eyes*, only the moving hands trigger events, whereas with *moving eyes*, events are generated by static as well as moving features in the scene, thus both static (*e.g.* the face) and moving objects (hands) appear salient. The rightmost column shows the x coordinates (along the axis between the person and the cameras) of peak saliency plotted against time (frame number) for both datasets. The true locations of the hands are marked with coloured bands. For *static cameras* (Row 2), the peak saliency pixel consistently alternates between the left and right hand locations as they move towards and away from the camera, *i.e.*, it follows the hand closest to the camera. For *moving eyes*, (Row 3), excess events caused by micro-saccades result in some spurious saliency peaks at objects like the face despite them being farther away from the camera.

by the attention models.

The direct comparison between the evProtoDepth and the fbProtoDepth qualitatively on *hands* dataset depicts that the fbProtoDepth shows a wider and centre-biased response, whereas the evProtoDepth shows a more localised response because event-driven cameras only respond to motion and high contrast changes and generate sparse features. The fbProtoDepth takes the entire human as single object due to the presence of additional

orientation and colour opponency channels, whereas in case of evProtoDepth, the event cameras produce sparse and disjointed features leading to the detection of multiple smaller objects. This can be observed in supplementary Fig A.10[2].

The evProtoDepth model is able to focus the attention towards the target which is closer to the robot, making it more suitable for behavioural decisions and interaction within its proximity. The system shows reliable response in cluttered scenarios and dynamic scenes. The Disparity Extractor alone provides a disparity map without any higher level filtering of "objects" in the scene. Therefore, the integration of the evProto model with the disparity extractor informs the system about salient regions which are not only nearby but also follow Gestalt laws. The proto-object model helps select a proto-object following Gestalt laws while discarding noise from the disparity map, whereas the additional disparity information improves selection precision in evProtoDepth. For evidence we point the reader to the supplementary Fig A.11[2], which depicts 2D histograms of peak responses for evProto, Disparity and evProtoDepth saliency maps.

4.7 Discussion

We introduce a model that combines disparity computations based on neuromorphic event-driven algorithms and hardware with a bio-inspired attention model. It improves upon the 2D model (the evProto model) which assigns perceptual saliency to (moving) edges that enclose a region (not necessarily completely) and can hence form the contour of an object. Adding the disparity information results in our 3D evProtoDepth model which, in addition to the salience imparted by evProto, assigns additional saliency to regions that are also closer to the cameras compared to those at larger distance. Adding depth information provides more stable object selection and robustness to noise, as demonstrated in Fig 4.5 and 4.6.

From the results presented about disparity estimation (see Fig 4.4 and Supplementary Fig A.7)[2], the event-based disparity estimation is robust and reliable in different scenarios with dynamic objects of increasing complexity. It can solve the correspondence problem for multiple objects simultaneously, distinguishing their relative distance from the robot. When the stream of events increases because of clutter and/or eyes movements, the accuracy of the disparity estimation is traded-off with latency, increasing the level of noise. Typically, the disparity information successfully enables the attentive system to select the nearest proto-object. The online evaluation implemented on a robot using real-world data proves the capabilities of the model in a realistic scenario. The system is robust to clutter and it demonstrates robust selection of the nearest proto-object in a noisy background.

The robot is responsive to motion, giving preference to closer moving objects. When we enable eyes motion, it can also select the nearby static object. The model tolerates motion of the cameras and of scene objects and usually determines as salient those areas that are closest to the cameras. The use of a biologically inspired event-driven disparity extractor distinguishes the evProtoDepth model from its frame-based counterpart fbProtoDepth. While the latter requires a pre-computed depth map from RGB-D sensor and computes feature maps representing local intensity, colour opponency and orientations, the only input required by our new evProtoDepth are raw streams of events from two neuromorphic cameras. Disparity information is extracted directly from these event streams using a bio-inspired cooperative matching algorithm. Benchmarking on the NUS3D dataset shows that despite those differences both models achieve similar performance, with the event-driven one being more easily applicable to online robotic applications, thanks to a more localised response over the selected objects.

Both models, fbProtoDepth and evProtoDepth, have strict bottom-up (data-driven) architectures and achieve mediocre results on the MIT metrics when directly compared with the eye fixation maps. This is expected due to the presence of complex attention mechanisms which include influences that are not captured by either of the models. These influences include cognitive top-down (goal driven) mechanisms, previous stimuli or priming [141] among others. As such, the quantitative comparison with the ground truth fixations of the NUS3D dataset, needed for a formal evaluation of the model, does not capture the system's true merit, that is, the robust selection of nearby objects in dynamic environments within 170 ms.

Although the saliency maps from the model can thus not be directly compared with fixation maps, the model still reasonably represents interesting regions of the scene. In general, the evProtoDepth model shows a more localised response to near-objects when compared with the fbProtoDepth model (see Fig 4.3). We believe this is mainly due to the sequential nature of processing in the event-based model. The simulated events used in this case first extract contrast information from the scene. Subsequently, only the depth information at event locations is used to inform the proto-object model. Therefore, the evProtoDepth model, having only one channel (depth), inherently prioritises the closer objects. In contrast, the frame-based model combines information from multiple channels (depth, colour opponency, intensity, orientations) at the latter stage of the pipeline, causing multiple features to contribute to predict the salient regions. The combination of cues from multiple channels produces a more dispersed overall saliency response. This may also lead to the fbProtoDepth model selecting objects with high contrast edges possibly located far away from the camera. We believe that prioritisation of close objects, at the cost of decreased attention to distant objects, is of high importance for a robotic agent because of its

need for interactions with physical objects. Nevertheless, on the long run information from different sub-modalities and from different distances needs to be integrated and weighted appropriately. The proposed model acts as a first milestone towards more complex robotic attentive systems that can include other important cues such as contrast, motion, colour and orientation. Furthermore, in future developments, such an entirely data-driven system could be enriched with top-down mechanisms, enabling the machine to switch priorities between extracted features depending on the robot's behavioural goals.

Additionally, in a more complete robotic pipeline, the saliency map could drive the robot's gaze in a more natural way. In fact, humans continuously gaze in order to bring the region of interest onto the fovea. In another work [179], we proposed an eccentricity model for sub-sampling the input visual space similar to that performed by a biological retina. In brief, the periphery of the field of view has coarser resolution than the middle (fovea). Combining such a model with an attentive system could be used in a pipeline that exploits saliency to drive the robot's eyes towards the most interesting regions, thereby giving salient regions a higher sensory resolution required for higher-level processing. This mechanism would both bestow the robot with a natural behaviour similar to that found in biology, and would also lead to savings in computational resources, since only salient regions are processed at the full resolution.

This work attempts to bridge the gap between biologically plausible saliency models and bio-inspired hardware. We demonstrated the model running online on a humanoid robot in different scenarios proving how event-driven cameras are well-suited for saliency detection in embodied agents. Stereo event cameras allowed the easy extraction of moving edges, solving the correspondence problem using precise spiking times, and the removal of layers of processing from the fbProtoDepth. The long term goal would be to implement such a complex algorithm onto neuromorphic specialised platforms [130], [167] to better exploit the event-driven pipeline aiming to further decrease the computational cost of the system in terms of latency and power consumption.

4.8 Methods

Traditional frame-based cameras generate frames synchronously at a fixed rate regardless of changes in the scene. For this reason the output contains great amounts of redundant data, especially in case of static scenes. Unlike regular cameras, event-driven sensors overcome the data redundancy providing data-driven output. This is particularly suitable for online robotic applications [115]–[118] given the need for low latency and high speed [114], [148]. Event-driven cameras react to illumination changes at the pixel level,

generating an asynchronous stream of events. Each event is defined as a tuple (x, y, p, t), where x and y are the spatial coordinates of the instantaneously active pixel, p the polarity bit encoding the direction of the illumination change (dark-to-light or light-to-dark), and t timestamp when the event occurs at microsecond resolution. An example of an event stream plotted in spatio-temporal coordinates is shown in the middle column of Fig 4.1.

In this study, we combine evProto [166], a previously developed event-based model for attentional selection with fbProtoDepth, a frame-based proto-object model that incorporates depth information[88] to develop the first version of an event-driven based saliency model in 3D which we call evProtoDepth. The current model uses depth as the primary channel for computing saliency. Depth perception is introduced *via* scene disparity extracted from stereo event cameras. Disparity is extracted using an asynchronous event-based bio-inspired cooperative neural network able to solve the correspondence problem [11] in a scenario with multiple objects. The disparity-encoded events from the disparity extractor are accumulated into non-overlapping disparity frames of 100ms duration, and are processed by the Border Ownership and Grouping Pyramids mechanisms in evProto to form proto-objects in the disparity map. An overview of the processing pipeline of the evProtoDepth model is presented in Supplementary Fig A.1[2]. We designed and implemented the model for real-time usage on the iCub robot.

4.8.1 Event-driven disparity extraction

In robotics, depth cues are important to select reachable objects upon which the robot can act, in addition to providing input for other tasks. The fbProtoDepth model uses depth from an RGB-D sensor. In order to implement a fully bio-inspired pipeline, we use disparity estimation techniques using stereo event-driven cameras as input for the evProtoDepth model. Binocular disparity of a 3D point relays information about its distance from the plane of fixation, but suffers from the problem of false correspondences. It is now widely accepted that mammalian brains solve this problem relying on a competitive process in disparity-sensitive neuron populations to encode and detect horizontal disparity [203]. Neurons compete with each other to represent the disparity of the scene, by removing false matching to reach a global solution. In particular, a disparity Cooperative Network [204] employs correspondence between a stereo event-pair, and it imposes disparity uniqueness and continuity conditions to construct a map representing the level of belief/confidence of corresponding points.

Asynchronous cooperative matching processing is well-suited to exploit the output of event-driven cameras since the precise timing of event generation can be used to find cor-

respondences efficiently at pixel-level without the need for patch or feature-based matching. This can produce disparity maps that can adapt to a dynamic input scene in real-time. Event-based cooperative matching algorithms have been efficiently implemented on neuromorphic platforms using Spiking Neural Networks (SNN) [125], [205] as well as on traditional computing platforms[11], [206]. Although specialised neuromorphic hardware [130], [167] is well-adapted for spike-based computation due to its low latency and power consumption, these new generation devices have difficulty handling networks with hundreds of thousands of neurons working in real-time on robotic platforms which demand robustness. This model implements an array-based representation of a Spiking Neural Network (SNN) based on an Event-based Cooperative Stereo Matching [11], similar to the SNN proposed by Osswald [125]. Our work implemented a real-time version of this algorithm on a standard CPU, prioritising its ease of deployment on the iCub and integration with the proto-object model over power consumption and efficiency afforded by neuromorphic hardware. It uses a 3D voxel-grid in $x-y-d$ space (d=disparity), called an *activity map*, which is updated asynchronously with each incoming event. Each element (cell) of this array represents a computational neuron in the SNN, which spikes during simultaneous triggering of events in the left and right camera. To ensure that temporally close events have higher probability to correspond to each other, a simplified version of the Leaky Integrate and Fire (LIF) model[207] is used to model the internal dynamics of each activity cell.

The output disparity value d for each pixel corresponds to the layer with the highest activity (belief) for that pixel. Each incoming rectified event affects multiple cells in the *activity map* through excitatory and inhibitory connections. The excitatory connections enforce continuity constraints by ensuring that neighbouring pixels have similar disparity values, implementing the prior that most surfaces in the 3D environment are continuous and smooth. The inhibitory connections enforce uniqueness constraints by suppressing false correspondences between stereo-pairs along the line of sight. They ensure that each pixel is assigned only one disparity value. The strength of interaction is determined by the time difference between successive interactions, such that a cell affected by multiple events in close temporal proximity will be highly active. The activity generated by each incoming event on a particular voxel is inversely proportional to how far in the past that voxel was last affected. After several cycles of excitation and inhibition within the activity map, a *disparity event* is generated by the network by associating the incoming event with the disparity value of the layer that has the highest activity. The output of the network consists of estimates of the disparities of all events and collects them in a single channel of disparity events E_d in the reference view of the left camera frame. With the event-based cooperative matching algorithm, we gain improvements over frame-based processing algorithms in

terms of processing time at the cost of accuracy of disparity. The resulting disparity maps are sparse and prone to noise, especially when the input event throughput is high, *e.g.* when the camera moves in a textured scene. However, this suits our needs as the downstream proto-object saliency model acts as a filter that suppresses noise in the disparity maps while selecting the nearest object (*e.g.* Fig 4.7). A schematic illustration of the network architecture is shown in Fig A.2[2]. Further details about the disparity extraction algorithm is provided in the supplementary material.

4.8.2 Proto-object based saliency with depth information

Variations and extensions of proto-object saliency models using frame-based cameras include the addition of addtional features including motion[87], texture [89] and depth[88], [109] (we call the latter fbProtoDepth). Each information channel is separately processed by a "grouping" layer, that represents proto-objects in the final saliency map combining all channels.

A previous event-driven implementation of the proto-object model [166] (evProto) focused on the use of event-driven cameras. The model exploits the inherent edge extraction capabilities of event-driven cameras, allowing it to omit the Gabor and center-surround filtering of the original frame-based model[9]. The output from the cameras is directly fed into the Border Ownership layer and processed in the same way as in the original version, detecting salient regions of the scene with a latency of $\approx$170ms every time there is a change in the scene.

The fbProtoDepth model[88] uses intensity, orientation, colour opponency and depth channels in parallel to compute saliency. In the evProtoDepth model, we implemented a single depth information channel, in the form of disparity-weighted event frames, fed into the grouping layer of the evProto model. The disparity of each individual event (based on the input from both cameras) is computed using a cooperative network model. Each output *disparity event* E_d contains information about the pixel (x, y), generation time ts and disparity estimate d of the corresponding visual stimulus. Disparity events arriving within a time window δt are accumulated in a disparity frame $D(x, y, t)$. Each of its pixels stores the disparity value d of the latest disparity event E_d emitted within that temporal window $(t - \delta t, t)$ at pixel (x, y). The length of the time-window is selected based on the desired sparseness of the disparity map fed into the grouping layer of the evProto model. The disparity frames are subsequently normalised within [0, 1] and passed onto the evProto model. While the input map in the original evProto model accounted for the presence of edges, our implementation extends this representation by also encoding the depth of each

edge. The key components of the evProto model[166] and proposed evProtoDepth models are shown in Supplementary Fig A.1[2].

Consent Statement

Informed consent has been obtained from the respective individual to publish images (Figs 4.7 and A.10) in an online open access publication.

Data Availability

The datasets generated and analysed during the current study are available from the corresponding authors on reasonable request.

4.9 Reflections & Conclusions

This implementation solves some of the problems of the intensity channel introduced in Chapter 2. It is robust to clutter and most importantly it focuses steadily on a specific target, namely the closest one.
The proposed work takes inspiration from the RGB version of the 3D proto-object model running on MATLAB [88].
The event-driven implementation allows a proper reaction of the robot to external stimuli. The system runs online exploiting the C++ speed for the disparity extractor and the Python implementation to generate the saliency map. The full pipeline takes 170ms to produce a stable saliency map, 100ms for the disparity map generation and 70ms for the proto-object detection. Regardless of all of these remarkable results the model still takes advantage of the event-frame representation feeding the proto-object model with a disparity map. The architecture would benefit if implemented fully on a neuromorphic platform such as SpiNNaker decreasing from 70 to 16ms the proto-object detection. Moreover, the disparity extractor could be implemented on SpiNNker [125], [133] further decreasing the time to obtain the disparity map.
The results show a prioritisation of the proto-object detection over closer non-proto-object items. Therefore, confirming the need for proto-object processing to avoid the detection of close noise or clutter. The pipeline voluntarily prioritises the closest object to the robot towards a possible physical interaction with it. The definition of this clear task would finally define a loss function where the robot needs to reach the closest object in the scene. The fixations maps from human subjects would still be a validation of the model and not

the ground truth. Bottom-up and top-down mechanisms would still work together even if the task has been defined. In order to obtain a more fair comparison with the fixation maps, the model would benefit from the addition of a centre bias mechanism [208].
One of the main achievements of this work is the increased robustness of the salient point (see the video: https://zenodo.org/record/5091539). The video clearly shows the difference between the evProto and the evProtoDepth when a multi-object scenario is happening. Along with the increased robustness, another great achievement is the confirmation of the model not being driven by the number of events but still prioritising close proto-objects. The sparse disparity maps are enough representative of the scene to allow proto-objects detection. Although the claim of the proposed model is the detection of complex stimuli like the human body, a detailed investigation would be necessary. Also, a more in-depth investigation into the accuracy of the disparity estimation would be necessary. All the experiments see scenarios where the background is mostly plain and not enough cluttered for the strong claim made. Experiments should further take into account fast dynamic scenes and clutter scenarios.

Chapter 5

Event-based eccentric motion detection exploiting time difference encoding

The supplementary materials can be found in Appendix B section B.

This work has been published.

D'Angelo, G., Janotte, E., Schoepe, T., O'Keeffe, J., Milde, M. B., Chicca, E., & Bartolozzi, C. (2020). Event-based eccentric motion detection exploiting time difference encoding. Frontiers in neuroscience, 14, 451.

5.1 Personal Contribution

Theoretical Implementation	Code development	Experiments design	Experiments execution	Analysis of the results	Paper writing	Academic authorship
yes	yes	yes	yes	yes	yes	First author

5.2 Authors

Giulia D'Angelo, Corresponding First Author, Istituto Italiano di Tecnologia - The University of Manchester
Ella Janotte, Technical Faculty, Bielefeld
Thorben Schoepe, Technical Faculty, Bielefeld
James O'Keeffe, The University of Newcastle
Moritz B. Milde, Western Sydney University

Elisabetta Chicca, Technical Faculty, Bielefeld
Chiara Bartolozzi, Istituto Italiano di Tecnologia

5.3 Authors Contribution

C.B. and G.D conceived the main idea behind the work with the help of J.O. G.D. developed the theory and the code for the eccentric down-sampling. G.D developed the entire pipeline integrating the down-sampling with the sEMD with help from E.J. G.D. designed the experiments with supervision from C.B. and M.M. G.D. conducted experiments with help from E.J. G.D. analysed the experimental results with help from E.J. G.D., E.J, T.S wrote the manuscript with the supervision of C.B, E.C, M.M. J.O. gave valuable feedback and helped edit the manuscript.

5.4 Abstract

Attentional selectivity tends to follow events considered as interesting stimuli. Indeed, the motion of visual stimuli present in the environment attract our attention and allow us to react and interact with our surroundings. Extracting relevant motion information from the environment presents a challenge with regards to the high information content of the visual input. In this work we propose a novel integration between an eccentric down-sampling of the visual field, taking inspiration from the varying size of receptive fields (RFs) in the mammalian retina, and the Spiking Elementary Motion Detector (sEMD) model. We characterise the system functionality with simulated data and real world data collected with bio-inspired event driven cameras, successfully implementing motion detection along the four cardinal directions and diagonally.

Keywords: Attentional Selectivity, Motion Detection, Eccentric Down-Sampling, Spiking Elementary Motion Detection, Bio-Inspired Visual System, Humanoid Robotics, Event Driven

5.5 Introduction

Most modern robotic systems still lack the ability to effectively and autonomously interact with their environment using visual information. Key requirements to achieve this ability are efficient sensory data acquisition and intelligent data processing. Useful information about the environment (e.g., how far away an object of interest is, how big it is, whether it is moving) can be extracted from sensory data. More complex interactions, for example locating and retrieving a particular resource, require an attentive system that allows robots to isolate their target(s) within their environment as well as process complex top-down information.

There are a number of ways for autonomous robots and natural organisms alike to gather information about their surroundings. Teleceptive sensors, for example those using ultrasound or infra-red light, are common in engineered systems, and are also exploited by some natural organisms for navigation and object tracking [209]. However, a closer relationship between attention and activation in the visual cortex has been observed by [210], showing the importance of vision when interacting and being attentive within an environment whilst performing a task. Motion detection, in particular, represents one of the important attentional cues for facilitating agent-environment interactions [51], and is used by natural organisms to avoid obstacles, respond quickly and coherently to an external stimulus within a scene, or to focus attention to a certain feature of a scene [52]. Due to its wide range of applications, motion detection has been an area of research for decades and has produced a number of different detection models, ranging from gradient-based algorithms [211], [212], over local-plane fitting [213], [214] and time-to-travel methods [144] to correlation-based approaches [215]. Gradient-based methods utilise the relationship between the velocity and the ratio between the temporal and the spatial derivative. Hence, to determine the speed and direction of the motion, the derivation of the spatial and temporal intensity for each pixel is needed. All correlation-based models share the linear and spatio-temporal filtering of measured intensities, which are functions of time and location. The best-known correlation motion detectors are the biologically derived Hassenstein-Reichardt and the Barlow-Levick models [142], [143]. The Hassenstein-Reichardt model was derived from behavioural experiments with beetles, while the Barlow-Levick model was inspired by motion detection in the rabbit's retina. In both cases one elementary motion detection unit is selective to motion in one cardinal direction (preferred direction) and suppresses output to motion in the opposite direction (anti-preferred direction) [143]. The models themselves (from 1956 and 1964, respectively), are still assumed to describe motion detection in organisms such as fruit flies [216]–[220]. A limitation of correlation-based detectors is that, depending on the time-constant of the filters used, the detector is

only receptive to a limited range of velocities. This range can be shifted by varying the parameters but always remains limited.

Environment analysis using traditional frame-by-frame visual processing generally requires a robot to extract and evaluate huge amounts of information from the scene, much of which may be redundant, which hinders the real-time response of the robot. The computational resources required for visual processing can be significantly reduced by using bio-inspired event-based cameras [5], [113], where the change in temporal contrast triggers asynchronous events. Event-based cameras perceive only the parts of a scene which are moving relative to themselves[1]. Thus, they are idle until they detect a change in light intensity above a relative threshold. When this happens, the pixel reacts by producing an event characterised by its time of occurrence. Address Event Representation (AER) protocol allows the asynchronous readout of active pixels while providing information on the the event polarity and the pixel location. As such, the camera's output are ON-events for increments in temporal contrast and OFF-events for decrements. Optical flow, the vector representation of the relative velocity in a scene, has a wide range of uses, from navigation [214], [221], to predicting the motion of objects [222]. We propose that these models can also be used to direct attention towards moving objects within a scene. Recent studies have developed event-based motion detection for optical flow estimation both relying on conventional processing archtitectures [120], [212], [223]–[225] and unconventional neuromorphic processing architectures [10], [226], [227]. Even though the former mechanisms, which leverage standard processing capabilities, show real-time optic flow estimation with very high accuracy, they are not suited for spiking neural networks and neuromorphic processors. This is due to the way information is represented, using real values in these algorithms. Additionally, the power consumption and computational complexity in [120], [224] is too high for constrained robotic tasks. The neuromorphic approaches on the other hand can naturally interact with spiking networks implemented on low-power neuromorphic processing architectures as information is encoded using events.

In the last decade a number of spike-based correlation motion detectors have been introduced [10], [226]. Of particular interest to this work is the spiking elementary motion detector (sEMD) proposed by [10]. The sEMD encodes the time-to-travel across the visual field as a number of spikes (where time-to-travel is inversely proportional to velocity). The sEMD's functionality has been evaluated in Brian 2 simulations and on SpiNNaker using real-world data recorded with the Dynamic Vision Sensor (DVS) [10], [228]. Furthermore, the model has been implemented on a neuromorphic analog CMOS chip and tested successfully [10]. The implementation on chip presents a low latency and low en-

[1]The perception of events is based on intensity changes in the scene. These changes can be due to negative or positive temporal contrast change occurring not only from relative motion.

ergy estimate of locally occurring motion. It further offers the advantage of a wider range of encoded speeds as compared to the Hassenstein-Reichardt model, and it can be tuned to different working ranges in sympathy with the desired output. Event-driven cameras, compared with classic frame-based cameras, dramatically reduce the computational cost in processing data, however they produce a considerable amount of output events due to ego-motion. Previous implementations of the sEMD have applied a uniform down-sampling across the camera's visual field. However, recent studies have found that motion detection performance depends strongly on the location of the stimulus on the retina, due to the non-uniform distribution of photoreceptors throughout the mammalian retina [229]. Rod and cone density in the mammalian retina is high at the fovea, and decreases towards the periphery. The non-uniform distribution of photoreceptors in the retina has a strong role in speed discrimination, and it should be taken into account as an important factor in motion estimation. Taking inspiration from the mammalian visual system [230], [231], where Receptive Fields (RFs) linearly decrease in size going from the retinal periphery towards the fovea [232], we propose an *eccentric*, space-variant, down-sampling as an efficient strategy to further decrease computational load without hindering performances. A good approximation of the mammalian space-variant down-sampling is the log-polar mapping, describing each point in the 2D space as logarithm of the distance from the centre and angle. Given its formalised geometrical distribution, the log-polar mapping provides algorithmic simplification and computational advantages, for example for tasks such as moving a robot's cameras towards a desired vergence configuration [233], or binocular tracking [234]. Recently, the log-polar approach has been studied also for event-driven cameras, with the proposal of the Distribution Aware Retinal Transform (DART) [235]. Although the log-polar representation would better suit the implementation of the eccentric down-sampling, the results in polar dimension would not be comparable with the classic down-sampling of the sEMD with Cartesian coordinates. For benchmarking purposes, in this paper we use an approximate implementation of the mammalian space-variant resolution, based on Cartesian coordinates.

In this work, we propose a novel approach to spiking elementary motion detection, exploiting the non-uniform retina model as a down-sampling of the visual field. By combining the sEMD with eccentric down-sampling, this work aims to improve the computational efficiency of the motion computation and take a step towards a bio-inspired attention model where information at the centre of the field of view is of higher resolution and more heavily weighted than information at the periphery, allowing robots to exploit visual information to effectively interact with their environments in real time. The proposed architecture is suitable for simulation on neuromorphic platforms such as SpiNNaker [6], and offers the possibility to be easily implemented for recorded and live input data. To the

authors' knowledge, artificial motion detectors with eccentric filtering of the visual field are a novel approach to motion detection. Link to the authors' repository containing the model and the data: *https://github.com/event-driven-robotics/sEMD-iCub.git*

5.6 Methodology

The proposed work integrates bio-inspired eccentric down-sampling with the sEMD [10]. Our aim is to further decrease the computational resources required, by filtering the number of incoming events into the visual field, while maintaining a fine resolution in the centre of the visual field.

5.6.1 Eccentric down-sampling

Several physiological studies have explored the mammalian retina topography such as the blind spot, fovea and eccentricities [236], showing that receptive fields are uniformly overlapped in the mammalian retina [237]. The proposed eccentric down-sampling approximates the two-dimensional circular retina onto a square, maintaining a quadrilateral camera resolution (Figure 5.1b), where each RF spatio-temporally integrates the information within its area of sensitivity. The RF size of the squared approximation decreases linearly toward the foveal region, where each RF is defined by one pixel. All RFs of the same size create a square ring around the foveal region, with each successive ring framing the previous one. The eccentric down-sampling reproduces the RF overlap between RFs of consecutive rings ensuring the robustness in response all over the retina. However, the proposed model does not include the central blind spot present in mammalian retina.

Equation (5.1) and (5.2) describe the relationship between the receptive field size (R^s) and its distance from the foveal region, where (R^c_i) is the centre of the top left RF of each squared ring and $i = [1, .., n]$ is the number of squared rings over the retinal layer. The term x in Equation (5.1) represents the x axis of the camera where the origin is placed in the top left corner, $max[\mathbf{R}^s]$ is the maximum kernel size of the outermost peripheral ring, and d_{fovea} is the total distance from the periphery to the edge of the fovea.

$$\mathbf{R}^s(x) = -\frac{max[\mathbf{R}^s]}{d_{fovea}}x + max[\mathbf{R}^s] \tag{5.1}$$

$$\mathbf{R}^c_i = R^c_{i-1} + \frac{R^c_{i-1}}{2} \tag{5.2}$$

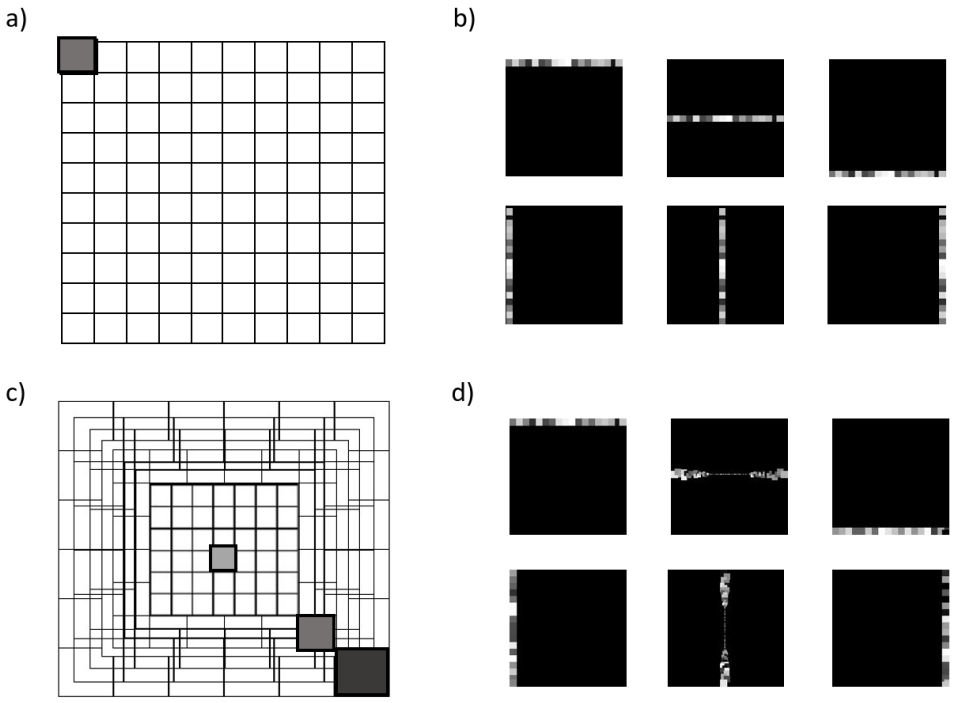


Figure 5.1. The grid in a) represents the uniform down-sampling of the visual field in equal matrices of n by n. c) Represents the eccentric down-sampling decreasing the size of the matrices going to the center of the visual field (fovea).This implementation does not include the blind spot present in the mammalian visual system. The three gray squares with varied hues represent three RF sizes at different eccentricities: 0, 39, 70 pixels distant from the centre. The square with the same hue in both grids (a, c) represents a matrix with equal size in the two down-samplings. Panels (b, d) represent the encoding in horizontal and vertical trajectories of the uniform down-sampling (b) and the eccentric down-sampling (d). On both top rows of (b, d), an example of the RFs belonging to the first, middle and last horizontal trajectories, and on the bottom row the vertical trajectories is given. All RFs are represented with different gray-scale for the reason of visualisation.

$$\mathbf{M}_t = M_{t-1} e^{-\frac{dt}{\tau}} + \frac{1}{R_{nf}} \tag{5.3}$$

Each RF is a matrix of input pixels from the sensor. Every RF is modelled as a leaky integrate and fire (LIF) neuron integrating the information in space and time (Equation 5.3), where M is the membrane potential of the RF, t represents the temporal information of the incoming event into the RF, dt the difference in time with the previous event in the RF, and τ is the time constant of the exponential decay ($\tau = 1000ms$). The membrane potential of every RF integrates incoming spikes until it reaches the threshold ($threshold = 1$), which is the same for all RFs. The contribution of each event to the increase in membrane potential of a neuron is normalised with the dimension of the RF. As the activity of the ATIS[2] is sparse, the normalisation factor (R_{nf}) is expressed as a percentage of the area of the RF. Every incoming event triggers the updating of the membrane potential by calculating the

[2]Event-driven cameras [5] used for this project.

temporal decay of the membrane since the last event. In addition, the membrane potential is increased by the normalisation factor. This way, the response from all RFs is normalised by their occupied space over the visual field. Finally, if the threshold is reached, the neuron emits an output spike. Hence, the response from each RF coherently encodes the input information in relationship with the distance from the fovea.

5.6.2 The spiking Elementary Motion Detector (sEMD)

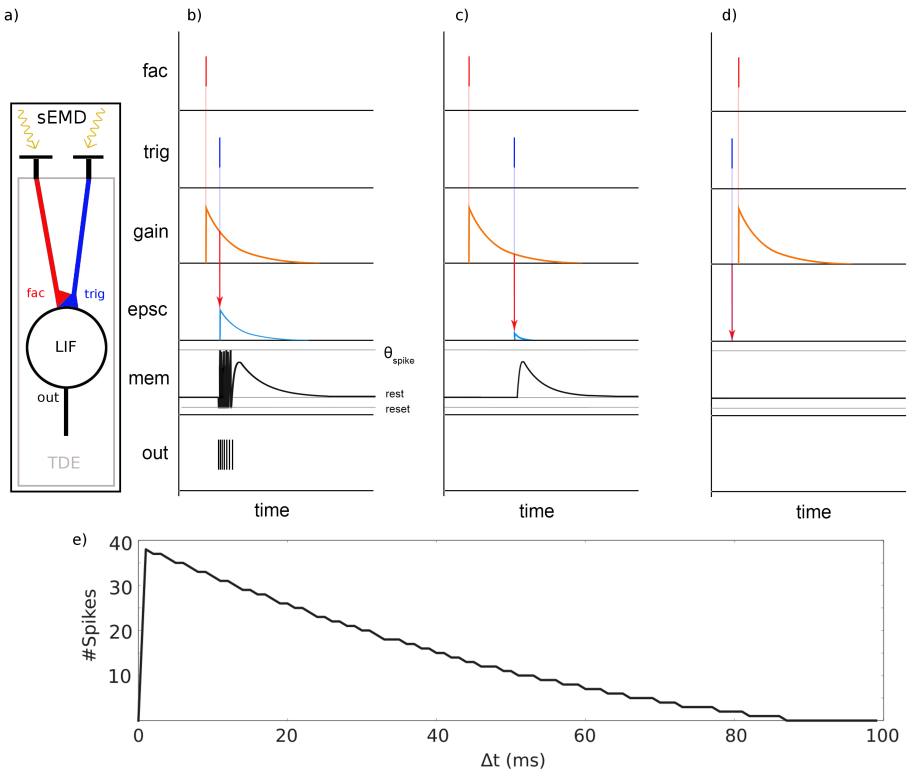


Figure 5.2. Basic principle of the sEMD [10]. a) The model consists of an event-based retina sending events into the Time Difference Encoder (TDE). Two adjacent RFs are connected to the facilitation synapse and the trigger synapse respectively. b) TDE computation for a small time difference Δ t between facilitation event and trigger event. An event at the facilitation synapse generates an exponentially decaying factor called gain. A trigger pulse at the trigger synapse shortly after causes an exponentially decaying Excitatory Post Synaptic Current (EPSC). The EPSC amplitude depends on the gain factor. The EPSC integrates onto the membrane potential (mem). Every time the membrane potential reaches the spiking threshold (τ_{Spike}) an output digital pulse is produced. c) Similar to b) part but with high Δ t. d) Similar to c) but the trigger pulse arrives before facilitation pulse. No output spikes are produced for negative time differences. e) TDE output spike response over time difference Δ t between facilitation event and trigger event.

The spiking Elementary Motion Detector (sEMD) depicted in Figure 5.2 has been designed for the purpose of encoding optic flow using event-based visual sensors [10]. The use of event-based sensors is suited to perceiving motion. The edge of an object moving from the receptive field of one pixel to the adjacent one generates a spike in the

two pixels with a given time difference, depending on the velocity of the edge and its distance from the pixels. The relative motion or optic flow is inversely proportional to this time-to-travel. An sEMD is composed of two pixels and a time difference encoder (TDE). The TDE encodes the time difference between two pulses into the number of output spikes produced in response to the second input pulse. The number of output spikes encodes the motion flow of objects moving in front of the two pixels.

The synapses connecting the inputs to the TDE are of two types - facilitator and trigger (see Figure 5.2 fac and trig). The facilitator synapse gates the activity of the TDE neuron. The trigger synapse elicits a response from the TDE neuron only if its input event occurs after the event from the facilitator synapse (compare Figures 5.2 b & d). The output current of the trigger synapse increases the TDE neuron's membrane potential as shown in Figure 5.2c). The strength of the current depends on the exponentially decaying gain variable of the facilitator synapse. Therefore, the TDE not only detects the direction of motion but also encodes the velocity of the stimulus in the number of output spikes and time to first spike. The faster the stimulus propagates, the more spikes are produced by the TDE. In order to mitigate the noise present at the output of a silicon retina, a pre-processing filtering stage is used. It consist of neural spatio-temporal filters (SPTCs) used to detect correlated events. Two uniform neighbourhoods, of n by n pixels, are connected to a LIF neuron each. The neurons fire once only if within a specific time, defined by their time constant, 66 % of the pixels in the neighbourhood produce events. The proposed implementation exploits the eccentric down-sampling (Chapter 5.6.1) replacing the uniform filtering stage previously used with the sEMD model by [10].

5.6.3 Experiments

The objective of this work is to quantitatively and qualitatively characterise the output of the TDE population receiving input from the eccentricity filtering layer and to compare it with the TDE population receiving input from a uniform resolution filtering layer. This characterisation aims to demonstrate the advantages of our proposed model, namely a decrease in computational load whilst maintaining the ability to estimate the velocity of moving entities within the visual field. To this purpose we characterised and compared the model using moving bars with 1D and 2D motion. In the following, we will refer to the two different implementations as "sEMD with uniform down-sampling" and "sEMD with eccentric down-sampling". The characterisation of the proposed motion detection system (Figure 5.3) is achieved using simulated data. Furthermore, additional experiments are

undertaken using real input[3] collected with ATIS cameras [5] mounted on the iCub robot (see Supplementary Materials for real-world data). The simulated data used in this work reproduces the activity of an event driven sensor in response to a bar moving horizontally (Left to Right (LR), Right to Left (RL)), vertically (Top to Bottom (TB), Bottom to Top (BT)) and transversely, i.e. along the diagonal of the Cartesian plane.

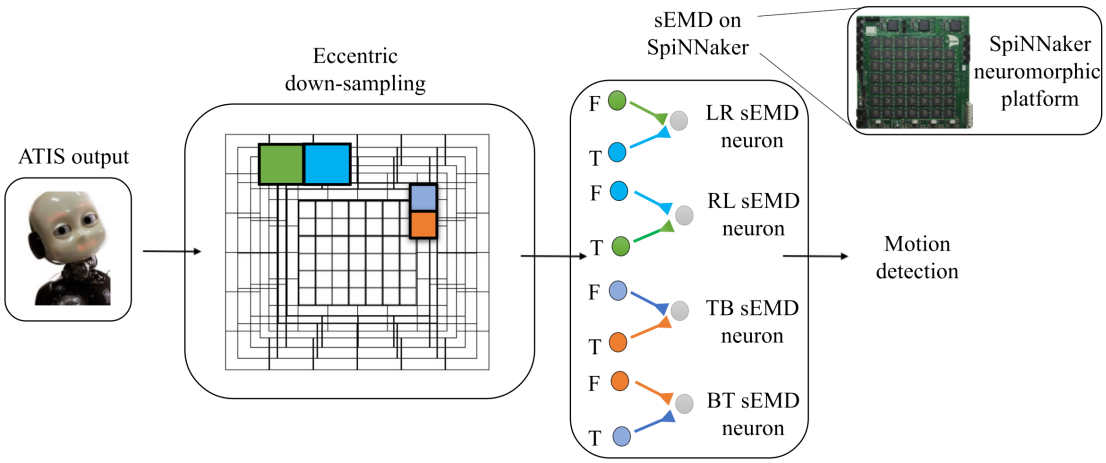


Figure 5.3. Basic scheme of the pipeline. From left to right the ATIS output is processed by the eccentric down-sampling model and sent to the sEMD model, hosted on SpiNNaker neuromorphic hardware. The sEMD model represents the layer of neurons producing spikes and encoding the motion detection. The eccentric down-sampling and the sEMD model representation show the spatio-temporal filter neurons (green, blue, violet, and orange square), the facilitator and the trigger, both synaptically connected to the sEMD neuron. Facilitators (F) and triggers (T) are shown for LR sEMD neuron, RL sEMD neuron, TB sEMD neuron and BT sEMD neuron.

Firstly, we recorded the activity of the sEMD with uniform down-sampling and eccentric down-sampling model, while the speed of the input bar ranges from 0.01 px/ms to 1 px/ms, in accordance to the experiments of [226]. This ideal input allows a comparison of the two model's spike raster plots and mean population activities.
We first analysed the selectivity of all sEMDs tuned to the same movement direction, measuring the mean firing rate (MFR) of the whole population. Given the symmetrical connectivity of the sEMD neurons along the eccentric visual field, the responses from the population of LR, RL, TB and BT sEMD neurons are expected to be comparable, responding with a large MFR to a stimulus moving along their preferred direction and being unresponsive to a stimulus moving along their anti-preferred direction.

Further investigations focus on a single population and its response to its preferred stimulus direction (from left to right, or top to bottom), assuming transferable responses

[3] We explored the real-world applicability of the underlying motion detection mechanism prior to this work in which we demonstrated the functionality of the underlying given variable contrast and event-rates in natural environments. (Milde et al. 2015, Milde et al. 2018, Schoepe et al. 2019)

for the other directions.

A deeper understanding of the temporal response from the neurons was achieved by collecting the spike raster plots for nine speeds of the chosen range: (0.01, 0.03, 0.05, 0.07, 0.1, 0.3, 0.5, 0.7, 1 px/ms), respectively.

For each speed, we analysed the response of each sEMD in the population, mapping its MRF onto the Cartesian space and visualising spatial rather than temporal information. We analysed how the Mean Firing Rate (MFR) of each sEMD changes with speed and distance from the centre of the field of view. Additional experiments have been performed changing the length of the stimulus, by recruiting more sEMDs, should increase the MFR of the whole population tuned to the corresponding stimulus direction. Eventually, we analysed the response of the model to a bar moving transversally exploring the response from the population to 2D motion. In such a case, the stimulus does not elicit the maximum response of any sEMD, rather, it elicits intermediate activity in more than one sEMD population, that need to be combined to decode the correct input direction.

5.6.4 Experimental setup

In all experiments the model was simulated on a SpiNNaker 5 board hosting 48 ARM-chips, each with 18 cores. The SpiNNaker architecture supports highly parallelized asynchronous simulation of large spiking neural networks in almost real-time. The aspect of real-time computation is of utmost importance for the interaction of the robot with the environment. For the implementation of the SNN we chose 160×160 pixels as a retinal layer resolution, to limit the number of neurons to be simulated on SpiNNaker and to further minimise the impact of the residual distortion in the fringes of the camera after calibration. The output of the retinal layer serves as input to the uniformly and eccentrically down-sampled filtering layer respectively. For the uniform down-sampling sEMD, we chose a non-overlapping neighbourhood matrix size of 4×4 ATIS pixels to represent one RF. This filtering layer is simulated on SpiNNaker and consists of 1600 LIF neurons. It receives input from a SpikeSourceArray, containing the respective ATIS pixel spike times. The synaptic weight of the connections is 0.3. In contrast, the fovea (1 RF = 1 pixel) of the eccentric down-sampling covers 10% of the total retinal layer, and the biggest receptive field has a dimension of 10×10 pixels with a normalisation factor of 60% (Equation 5.3). The population is made up of 8836 LIF neurons. The eccentric down-sampling occurs locally before the spike times of the respective receptive fields are transferred to SpiNNaker in a SpikeSourceArray. The final layer of the network consists of four sEMD populations sensitive to local motion in one cardinal directions respectively, using sEMD neuron model

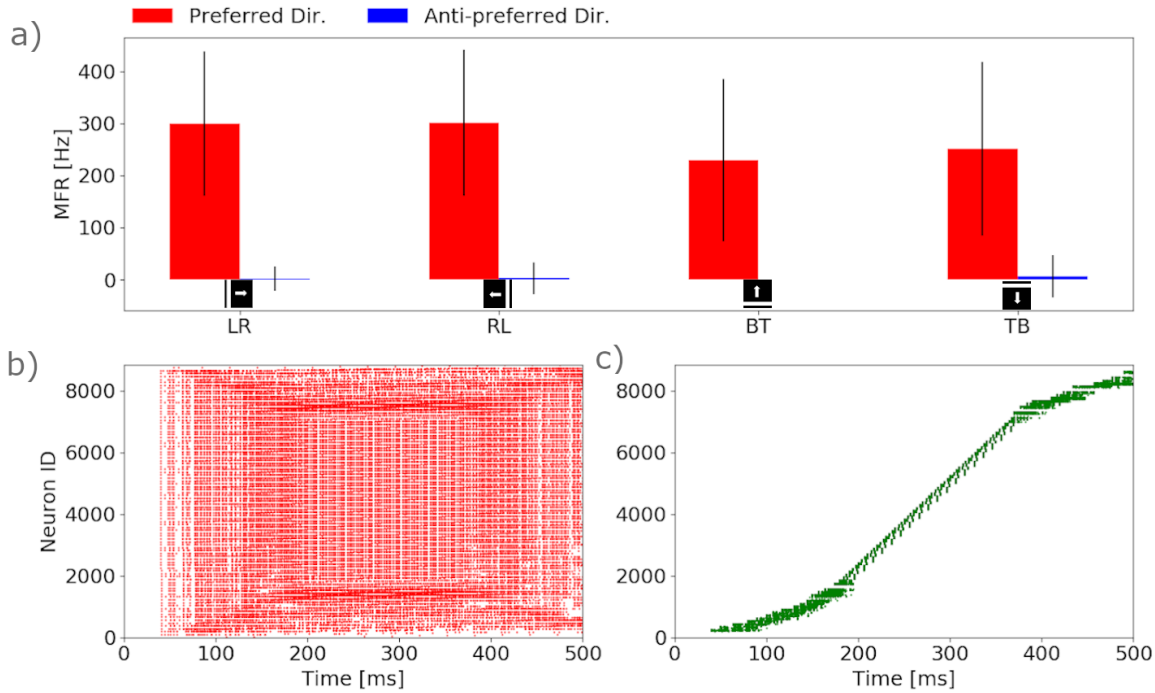


Figure 5.4. Response of the sEMDs with eccentric down-sampling to a simulated bar moving with a speed of 0.3 px/ms: a) Instantaneous MFR and variance of the four sEMD-populations, each tuned to one of the four cardinal directions, to the preferred and anti-preferred stimulus. Similarly to the sEMD with uniform down-sampling, the response to the anti-preferred stimulus is negligible with respect to the response to the preferred direction stimulus. b) Raster plot of the left to right (LR) population in response to a vertical bar moving from left to right. In the first 100 ms, the difference in the size of the RF can be seen, as the active neurons spike with different spike rates and the number of active neurons increases with time, when the bar moves closer to the fovea. c) Raster plot of the top to bottom (TB) population in response to an horizontal bar moving from top to bottom. The sigmoidal shape arises from the geometry of the eccentric down-sampling and the neurons' indexing.

included in the extra models of the pyNN library. The sEMD populations were connected to the filtering layers along the trajectories as shown in Figure 5.3. The combination of the output of the four populations allows the encoding of transversal stimuli. Each population shares the size of the down-sampling population. For both down-sampling approaches all sEMD neuron and synapse parameters are the same. The connectivity of the respective sEMD populations are displayed in Figure 5.3. The synaptic weights are 0.3 and the synaptic time-constants τ_{ex} and τ_{in} are both 20 ms. The neuron parameters amount to: a membrane capacitance of 0.25 nF, and time-constants τ_m and τ_{rf} of 10 ms and 1 ms respectively. The reset, resting and threshold voltage of the neurons are defined as -85, -60, and -50 mv respectively. To avoid a response of the sEMD-populations perpendicular to the preferred direction, in case of a bar moving their facilitator and trigger synapses receive input at the same time, the input to the facilitator synapse was delayed by 1 ms.

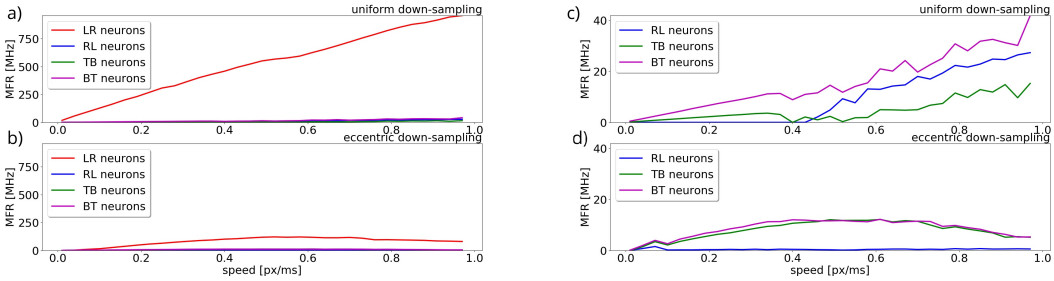


Figure 5.5. Comparison of the sEMD model with the uniform down-sampling (a, c) and the eccentric down-sampling (b, d) in response (MFR) to a left to right moving bar (simulated data). The preferred direction is displayed in red (LR), with the anti-preferred direction in blue (RL). The response for the top to bottom (TB) and bottom to top (BT) populations are displayed in green and magenta respectively. Panel (c,d) are a magnification for the anti-preferred direction (right to left) and the incorrect directions (top to bottom and bottom to top) of the panels (a,b). The Figure compares the behaviour from the populations of the two approaches to the same stimulus and over the same range of speeds.

5.7 Results

Our investigation starts with the characterisation of the eccentric down-sampling sEMD's response to a simulated bar moving in the four cardinal directions with a speed of 0.3 px/ms: left to right, right to left, top to bottom and bottom to top. Figure 5.4 shows the response to stimuli moving in the preferred and anti-preferred directions at fixed velocity 0.3 px/ms (the middle of the regarded velocity range). In particular, Figure 5.4a) shows the mean instantaneous firing rates of the preferred and anti-preferred direction populations. The preferred directions are coloured in red and the anti-preferred directions in blue. As expected, the preferred direction population's response is significantly higher than the response of the anti-preferred direction population. Furthermore, as expected the response from all the populations to the respective preferred direction is similar in terms of instantaneous firing rate and mean firing rate, and comparable among each other, thus validating the assumption that the response to stimuli in the preferred direction is similar for all of the populations. Assuming a bar moving across the retina at a constant speed, the high variances in preferred and anti-preferred directions can be explained by the difference in receptive field sizes in our proposed model (see Figure 5.1). Depending on the stimulus speed, the size of the RF determines a period of time in which the stimulus moves over the RF. Thus, for the same stimulus speed, a peripheral RF takes more time to respond than one in the foveal region, leading to a different RF rings having a different sensitivity to stimulus speed. Only the RFs along the same squared ring have the same sensitivity to the same speed. If a bar is moving across the visual field at a certain speed, only neighbour RFs, that produce spikes able to trigger the TDE neurons, will detect the stimulus. Consequently, due to the varying RF sizes and varying speed sensitivities, the size of the RF relative to its neighbour affects the response of the TDE. This causes the visual field

to respond non-uniformly. Figure 5.4b),c) show examples of characteristic raster plots of the preferred direction populations, in response to a bar stimulus moving horizontally and vertically, respectively. The colour-coding indicates the direction sensitivity of the population: left to right (red) and top to bottom (green). The first response to the horizontal and vertical bar movement (Figures 5.4b,c)), is delayed by 40 ms. This is due to the stimulus taking 30 ms (speed of 0.3 px/ms) to travel over the first peripheral RF (10 x10 px), before reaching the RF connected to the trigger. In the first 50 ms of reaction to the stimulus, the resulting spike density is rather sparse, caused by a lower response from the peripheral RFs (sensitive to higher speeds). Conversely, from 150 to 400 ms, the time where the stimulus is expected to cross the fovea, the spike density is higher because the RFs at the fovea are of a size more suited to the stimuli velocity. The impact of the proposed model is more clearly visible in response to the vertically moving stimulus (Figure 5.4c).The mapping from the eccentric receptive fields to the neuron IDs transforms the time sequence of a vertical bar response to a sigmoid. By contrast, the output of the sEMD with uniform down-sampling resembles the shape of stairs, with each row activated after one another, spiking with the same rate. The non-uniform size of the RFs in our proposed model is again the cause for the different spike densities produced in response to the stimulus moving at constant velocity. In this experiment the sEMDs successfully encode the direction of the bar stimulus moving across the visual field in all the four cardinal directions, showing a negligible response to the anti-preferred direction. This therefore shows that the eccentric down-sampling preserves the ability of the sEMD populations to encode optic flow of moving stimuli.

A comparison of the MFR for all populations of the uniform down-sampling model and the eccentric down-sampling model in response to a simulated stimulus moving from left to right at different velocities is shown in Figure 5.5. The color-coding remains the same as in Figure 5.4(b,c), additionally the response of the populations selective to stimuli from right to left and bottom to top is depicted in blue and magenta respectively. Figure 5.5 a) shows the behaviour of the uniform down-sampling model, and Figure 5.5 b) depicts the behaviour of the eccentric down-sampling model. Both methods show a trend of increasing MFR until target velocity reaches 0.6 px/ms. While the response from the sEMD with uniform down-sampling keeps increasing after 0.6 px/ms, the firing rate of the population with eccentric down-sampling gradually reduces as the target velocity approaches 1.0 px/ms. The same trend can also be seen for targets moving in the anti-preferred direction. Figure 5.5 shows that, while the sEMD response of the anti-preferred (right to left) and the incorrect directions (top to bottom and bottom to top) of the uniform down-sampling model (5.5 c) linearly increases[4] until 1.0 px/ms, the output firing rate of the proposed

[4]The linear increment of the TB and BT response from the neurons is due to the events occurring for the moving vertical bar.

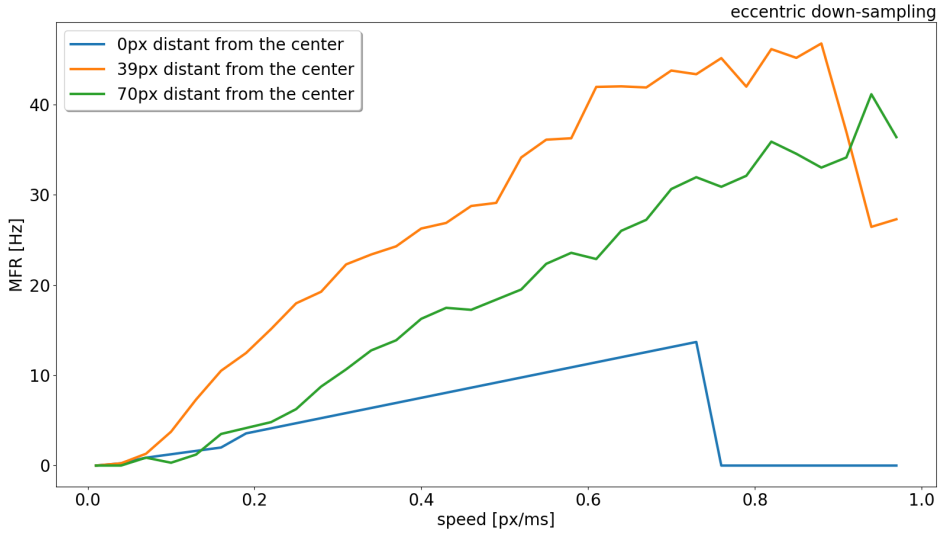


Figure 5.6. Response (MFR) to a left to right moving bar (simulated) from RFs (eccentric down-sampling) of the central horizontal line of the visual field at different eccentricities (distances from the center of the field of view). In blue, orange and green at 0, 39 and 70 pixels distant from the centre, respectively (see Figure 5.1).

eccentric down-sampling model (5.5 d) increases for target speeds up to 0.5 px/ms and decreases thereafter. Despite the number of sEMDs required for the proposed model (8836 per population) being significantly higher than for the uniform down-sampling (1600 per population) under the same setup conditions, the eccentric sEMDs' down-sampling shows an overall significant decrease in the mean output firing rate of the whole population i response to the same stimulus. Differently from frame-based systems, where the number of operations - and hence power consumption - depend on the number of filters, in event-driven spiking architectures, filters are active (and consume power) only when they receive input spikes and produce output spikes. Figure 5.5 shows that the proposed eccentric down-sampling model is able to differentiate between stimulus in preferred and anti-preferred directions more efficiently than a model with uniform down-sampling, without sacrificing performance. The proposed model still maintains an order of magnitude difference between MFR for stimulus in the preferred direction versus anti-preferred direction. Although the eccentric down-sampled model does not allow for an inference of stimulus velocity to be made based on the MFR of the entire population, the same information can be extracted based on the eccentricity of the RFs with the greatest MFR.

The response from sEMDs selected at different eccentricities (at 0, 39 and 70 pixels distant from the centre) is examined in Figure 5.6 in relation to the same speed range. In the original model [10] the MFR of all three neurons would increase proportionally to the

Events belonging to the same x but different y have a different timestamp.

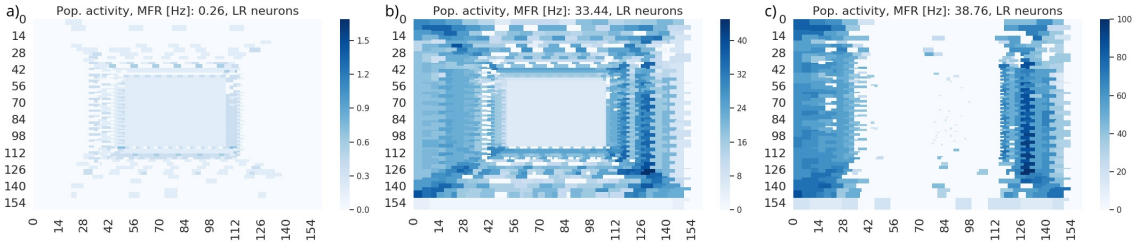


Figure 5.7. Response from the population of sEMDs with the eccentric down-sampling mapped into the cartesian space with a camera resolution of 160x160 pixels. The color-code heatmap represents the MFR of each RF. The stimulus was a bar moving (simulated data) from left to right with constant speed: 0.03 a), 0.3 b), 1.0 c) px/ms, respectively.

target speed. Figure 5.6 shows that the speed encoding for our proposed model depends on the RF size, because the integration time for each RF size corresponds to a specific range of velocities. This leads to a specific range of time-differences between two connected RFs. Each sEMD has a speed limit, which depends on its tuning, above which it will be unable to detect motion. Figure 5.2 e) shows the TDE output spikes over time difference. If a trigger event occurs before the output of the facilitation event has had time to reach the minimum threshold required, the sEMD will not fire. Due to the varying sensitivity of different RF sizes and enhanced by the 1 ms synaptic delay of the facilitator synapse, while the response from the foveal region (0 px distance) drops to zero for speeds higher than 0.7px/ms, the response from the neuron with a middle eccentricity (39 px distance) begins to decrease dramatically at 0.9px/ms. The response from the peripheral neuron keeps increasing until the end of the examined speed range (1.0 px/ms). A possible explanation for the relatively low MFR of the peripheral neuron is the increased number of events needed to trigger the RF and its specific sensitivity to higher speeds. Figure 5.6 shows how the RF size affects the behaviour of the correspondent neuron, obtaining a wider operative range from the whole population. In comparison, uniform down-sampling where all the RF sizes are the same provides a comparatively limited operative range.

The spike raster plots (Figures 5.4b,c) provide the temporal response from the population but they do not provide any spatial information. The visualisation in Figure 5.7 maps the response of the sEMDs to the corresponding x and y locations for three different speeds: slow (0.03 px/ms Figure 5.7a), medium (0.3 px/ms Figure 5.7b) and fast (1.0 px/ms Figure 5.7c). The data displayed in Figure 5.7b) corresponds to the spike raster plot in Figure 5.4b). Figure 5.7 shows that the MFR of the whole population increases in relation to the speed: 0.26, 33.44, 38.76 Hz, respectively. The spatial visualisation highlights the function of the eccentric down-sampling. As proposed by [229], the slow speeds are detected primarily in the foveal region, where RFs have the smallest dimension and are closest to one another (Figure 5.7a). As the stimulus speed increases, the peripheral re-

gion starts responding from the first squared ring around the foveal region (Figure 5.7b) to the rings with the largest RF size for the fast speed (Figure 5.7c).

The response for each RF square ring is different for horizontal and vertical components (most obvious example being in Figure 5.7c). This is because the sEMDs in this case are only connected horizontally (as we are working with left-right motion). Therefore, at the left and right peripheries, there is a descending and ascending scale of RF sizes approaching and moving away from the foveal region, respectively. A concentrated region of diverse, overlapping connected RFs improves the likelihood of the sEMDs picking up the stimulus motion. This does not exist in the regions above and below the fovea, in which each RF will only be connected to horizontally adjacent RFs of the same size, hence the relatively low MFR in these regions.

The response on the right side of the visual field is attenuated in Figure 5.7b and Figure 5.7 c because the sEMDs from the last RF ring are not connected with any subsequent facilitator (although this does not cause a problem in detecting stimuli entering the scene).

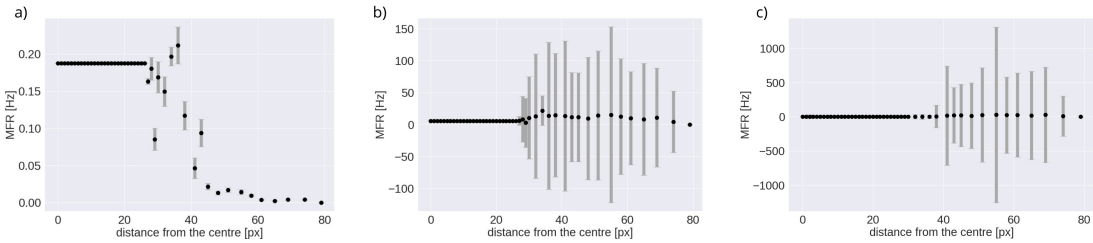


Figure 5.8. Mean and variance in MFR of RFs at different distances from the centre of the visual field. The stimulus is a moving bar (simulated data) going from left to right at speeds of: 0.03 (a), 0.3 (b), 1.0 px/ms (c).

As shown in Figure 5.7, the RF-ring of maximal response appears to move toward the periphery with increasing velocities. Figure 5.8 shows the mean and variance of the MFRs at different eccentricities for velocities 0.03, 0.3 and 1.0 px/ms, Figures 5.8a) (b,c), respectively. It is clearly distinguishable, that the maximal response in MFR shifts towards the periphery with increasing velocities.

The higher variances observed at greater eccentricities (distance from the centre) in Figure 5.8b) and c), can be explained by the different RFs response from the horizontal and vertical component of the squared rings (which can be seen in Figure 5.7). The low MFR at 29 pixels (Figure 5.8a) from the centre (fovea region from 0 to 28 px) can be explained by the connections between RFs of the first peripheral squared ring (about 3x3 px) and the fovea, where each RF has a dimension of 1 px. This sudden increase in size leads to a delay in response from the TDE receiving input to the trigger synapse from the larger receptive field.

To compare the trend of the RFs' peak response increasing in eccentricity with increasing stimulus speed, the center of mass of the RFs response is plotted in relation to the speed range, from 0.01-1.0 px/ms (see Figure 5.9). Figure 5.9 shows that for low speeds (0.01 to 0.06 px/ms) the centre of mass of the RFs' response shifts from 0 to 27 pixels (distance from the centre). The centre of mass then plateaus from 0.06 px/ms to 0.6 px/ms, where only the RFs of the edges of the foveal region respond to the stimulus. For higher speeds (from 0.6 to 1.0 px/ms), the eccentricity of the centre of mass of RF responses starts to increase again, due to a lack of activity in the fovea. The centre of mass of RF responses eventually shifts to the periphery, reaching a distance of 49 px from center.

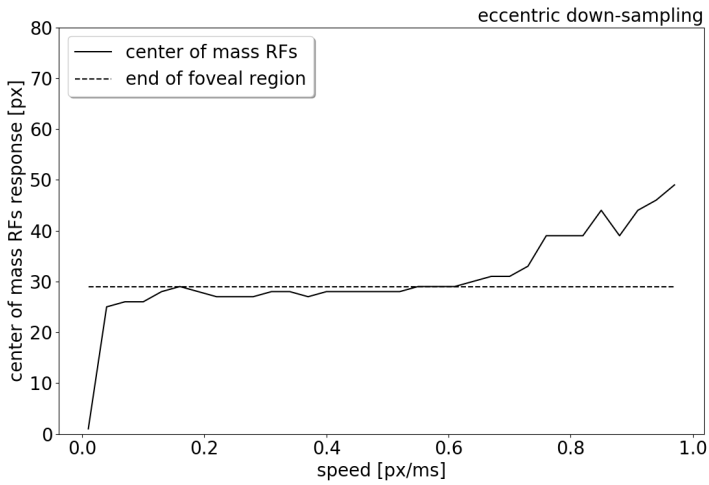


Figure 5.9. Center of mass (solid line) of the neurons response location to a left to right moving bar (simulated data), from 0.01 to 1.0 px/ms. The dash line indicates the end of the foveal region.

A comparison of the MFR of the sEMD with uniform down-sampling and eccentric down-sampling has been explored with simulated data. Figure 5.10 shows the difference in response, normalised for the total number of neurons, from all populations of sEMD neurons with uniform down-sampling and eccentric down-sampling. Even though the uniform down-sampling model has fewer neurons than the eccentric down-sampling model (1600 compared to 8836 neurons, respectively) the MFR from the eccentric down-sampling is considerably less at each explored speed, increasing computational and power efficiency.

Figure 5.11 shows the MFR from the population of LR sEMD neurons in response to a stimulus moving from left to right, at a medium speed of 0.3 px/ms, with bars of varying lengths: 10, 50, 100 and 160 pixels, respectively. The plot shows a positive correlation between the size of the bar and the response from the neurons sensitive to the corresponding direction. Figure 5.11 shows that the MFR increment decays as the length of the bar increases - most noticeable when comparing the difference in MFR between the 50 and

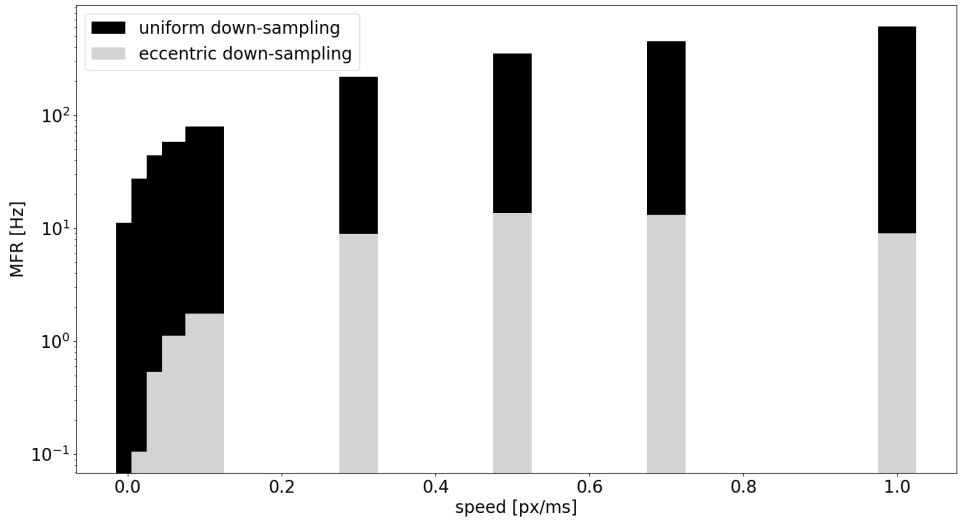


Figure 5.10. Comparison, between the sEMD model with the uniform down-sampling (1600 neurons) and the eccentric down-sampling (8836 neurons), of MFR from the LR sEMD neurons in response to a left to right moving bar.

100px bar, and that between the 100px and 160px bar. This is because the bar is vertically centred in the visual field, and so longer bars cover more of the peripheral region - where each RF requires a greater number of events in order to be activated. Finally, Figure 5.12 shows the behaviour of the population to a bar moving transversely, revealing the response of the model to 2D motion. Figure 5.12 a) shows the response to a bar moving from the top left corner to the bottom right, b) from the top right corner to the bottom left, c) from the bottom left to the top right corner and d) from the bottom right corner to the top left. All the explored cases report a similar response from two kind of sEMD populations and a response close to zero from the other neurons. The combination of the responding sEMD neurons successfully detects the transverse motion, showing similar MFR values of the neurons that actively respond.

5.8 Discussion

The biological role of detecting temporal changes comprise two mechanisms: the detection of fast and slow movements. The first one to identify an entering stimulus into the scene and the latter one to recognise its spatial structure [238]. Sudden onset of motion can attract our attention [52], [239]. Hence, fast movements, speed and acceleration similarly increase our perception of a threat - making it a noticeable stimulus and grabbing our attention [240]. Thus, motion detection collaborates with attentional mechanisms to react on

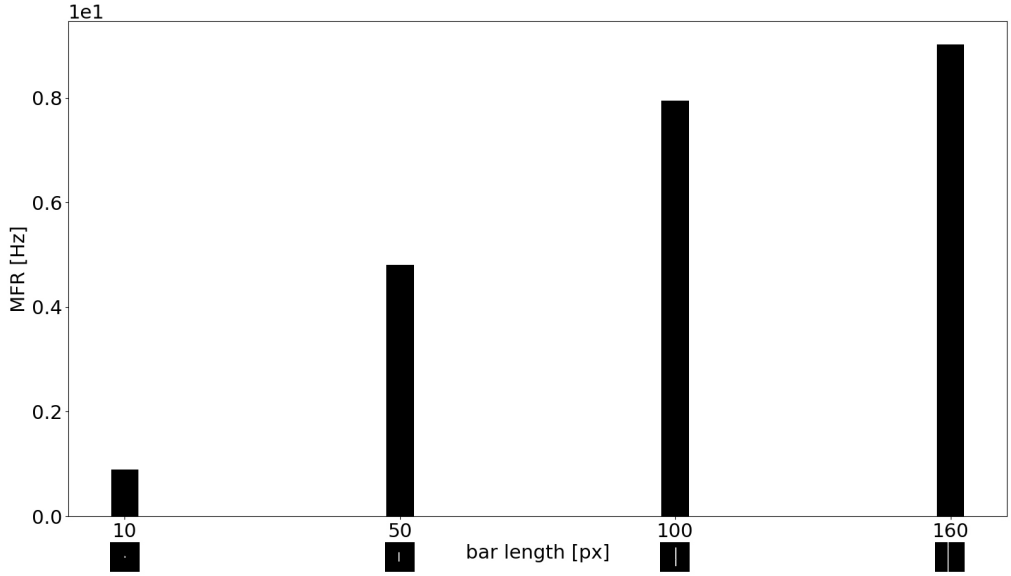


Figure 5.11. MFR response of the sEMD the LR sEMD neurons for a left to right moving bar at 0.3 px/ms with different bar lengths: 10, 50, 100, 160 pixels respectively.

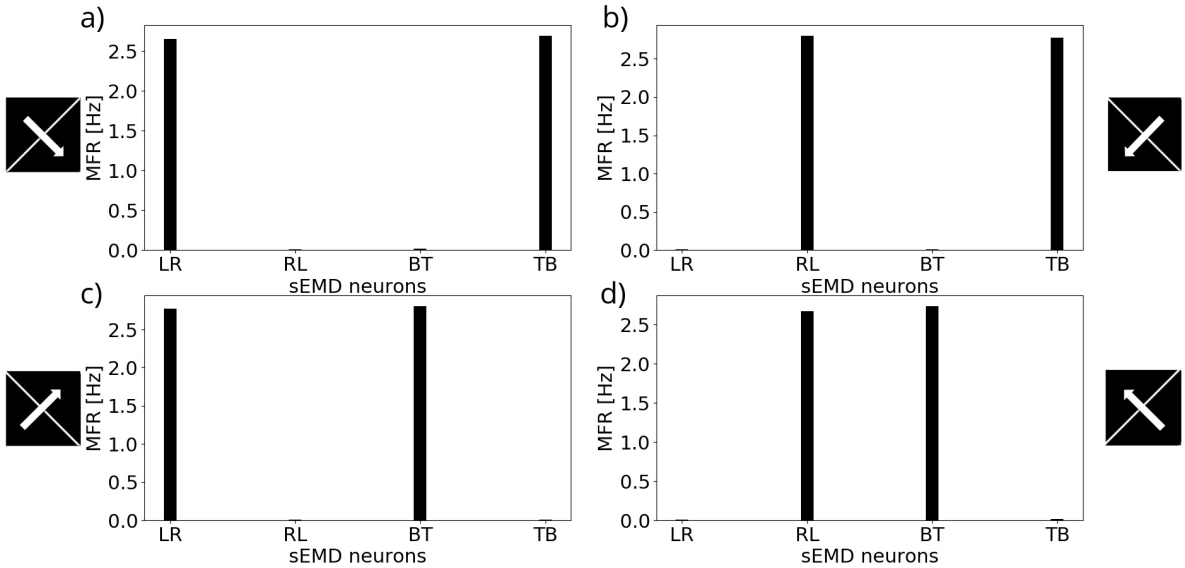


Figure 5.12. MFR response of the sEMD neurons reacting to a bar moving transversely at 0.3 px/ms. a) Bar moving from the top left corner to the bottom right corner, b) bar moving from the top right corner to the bottom left corner, c) bar moving from the bottom left corner to the top right corner and c) bar moving from the bottom right corner to the top left corner. The length of the bar covers the whole visual field.

time and interact with the surrounding. In this paper, we have presented a novel implementation of motion detection based on the use of spiking elementary motion detectors coupled with non-uniform down-sampling inspired by the mammalian retina. The proposed model successfully detects the correct direction of an edge moving in the field of view at

speeds ranging from 30 to 1000 px/s, being suitable for the coarse motion processing of robots interacting with the environment [226]. With respect to the uniform down-sampling implementation presented in the original work [10], the eccentricity model significantly decreases the overall activation of each motion detector at every investigated speed. The reduced spiking activity makes this implementation more power efficient even in face of an increased number of elementary motion detectors. To achieve the same result in the uniform down sampling implementation, the size of the spatio-temporal filters should be increased, at the cost of a coarser resolution in the whole visual field and a reduced sensitivity to low velocities. The eccentricity implementation overcomes this issue maintaining the sensitivity for low and fast speed — distributed over different regions of the field of view — while significantly reducing the number of incoming events to be processed by the down-stream computational layers.

In the proposed non-uniform down sampling, the elementary motion detectors are tuned to different ranges of speed depending on their position in the field of view. The peripheral sEMDs are characterised by large receptive fields and are hence tuned to higher speeds, that progressively decreases towards the fovea. Hence, the proposed implementation encodes the speed based on the location of the active sEMD. RFs with similar size work in a similar range of speed producing redundant information, and making the decoding of the population activity robust. Moreover, thanks to the sensitivity to high speeds of the peripheral RFs, the detection of objects moving into the visual field is immediate. The sEMDs in periphery will trigger a response to a fast stimulus entering the field of view with extremely low latency. This behaviour is desirable in our target scenario, where a robot shall react quickly to fast approaching objects suddenly entering the field of view, and attracting its attention. Furthermore, the combination of RFs with different size, processing events on the same field of vision, allows working with a wider operative range of speeds. In the final application[5], this motion detection module will be used as one of the feature maps used to compute the salience of inputs in the field of view, directing the attention of the robot to potentially relevant stimuli that will be further processed once a saccadic eye motion will place the salient region in the fovea. A strong and low latency response of peripheral sEMDs to fast stimuli could override the salience of static objects. The characterisation of the response of the sEMDs in the non-uniform down sampling shows the same qualitative overall behaviour for real-world stimuli, showing robustness to noise and to changing the overall spiking activity of the input. The analysis of the individual responses of the sEMDs at different distance from the fovea shows variability that depends on the discretisation of the receptive fields and on the uneven distribution of the receptive field sizes. This effect possibly depends on the Cartesian implementation of the

[5]The final application represents the full proposed PhD Thesis project.

eccentricity, that approximates the distribution of the receptive fields with a rectangular symmetry. A polar implementation of the same concept will reduce the effects of discretisation and improve the overall population response. In a polar implementation, the direction of each sEMD will be aligned along the polar coordinates (radius and tangent), rather than along the Cartesian directions, further improving the variability in the overall response of individual modules and allowing decoding of stimulus direction beyond the cardinal ones.

5.9 Data Availability Statement

The datasets generated for this study can be found in the https:// github.com/event-driven-robotics/sEMD-iCub.

5.10 Acknowledgements

We thank our colleagues from Italian Institute of Technology, Luca Gagliardi and Vadim Tikhanoff, who provided insight and expertise assisting the research. We would also like to show our gratitude to Jay Perrett for sharing his accurate review.

5.11 Reflections & Conclusions

The eccentric representation of the visual field further enhances the bioinspiration of the system, allowing better management of the events from the cameras. This implementation exploits a ”smart” subsample of the visual field increasing the size of the RFs in the periphery allowing a detailed vision in the fovea and coarse in the periphery. The system detects the motion direction thanks to two consequent RFs connected to a time-difference-encoding (TDE) neuron.
The big RFs in the periphery are sensitive to high speeds due to their bigger size with respect to the RFs in the fovea, providing a sense of alertness. The proposed model detects the speed and motion direction of a moving item granting a quick response from the robot for avoidance or interaction with the moving object. The model response depends on the fixed 1 ms delay between two consequent RFs (trigger and facilitator). We explored a speed range from 0.01 to 1.0 px/ms where the fastest speed was detected by the peripheral RFs, guaranteeing an immediate response from the system for an object entering the visual field.

The system runs on SpiNNaker, taking advantage of this neuromorphic platform to generate a fully spike-based pipeline. Furthermore, once the trigger and facilitator neurons are triggered, the response from the output neuron is immediate.
A further decrement of incorrect motion direction detection is visible and not appropriately explained in Figure 5.5. The uniform down-sampling shows a linear increment of not only the response of the anti-preferred direction but also of the TB and BT neurons. The eccentric down-sampling model further decreases this incorrect response thanks to the RFs size of the periphery. The events occurring (on the same x and different y values) do not trigger the "big" RFs. This implementation uses more neurons compared with the original version but they fire significantly less frequently to decode the same speed confirming a good administration of the input data. However, this achievement has not been investigated further by looking at the signal-to-noise ratio. Therefore, it would need additional experiments to confirm the efficiency improvement claimed. Nonetheless, the most important achievement is the wider speed range obtained thanks to the different sizes of the RFs. Although the proposed eccentric down-sampling widens the speed range detection, the visual representation in Figure 5.7 b) shows artefacts due to the activation of RFs with different sizes. Periphery and fovea are concurrently responding generating artefacts on the right part of the visual field due to the left-to-right motion triggering the RFs where x RF centre is between 126 and 140). This implementation could easily become log-polar detecting circular motion, incoming motion towards the fovea and the opposite direction adding another information for free: the angle motion direction over the retina (see Figure 5.13).
Motion seems to be an important cue in attention mechanisms modulation attention directly without the need to detect the item itself. Something moving is inherently interesting [52], and for this reason, this channel does not need to feed the proto-object model, contributing to the final saliency map together with the intensity and the depth channel.
The unfortunate issue of this model is the high sensitivity to noise making the non-preferred direction neurons slightly respond also for not ideal motion directions. This could be handled using lateral inhibition mechanisms all over the visual field, enhancing coherent motion among neighbouring RFs.

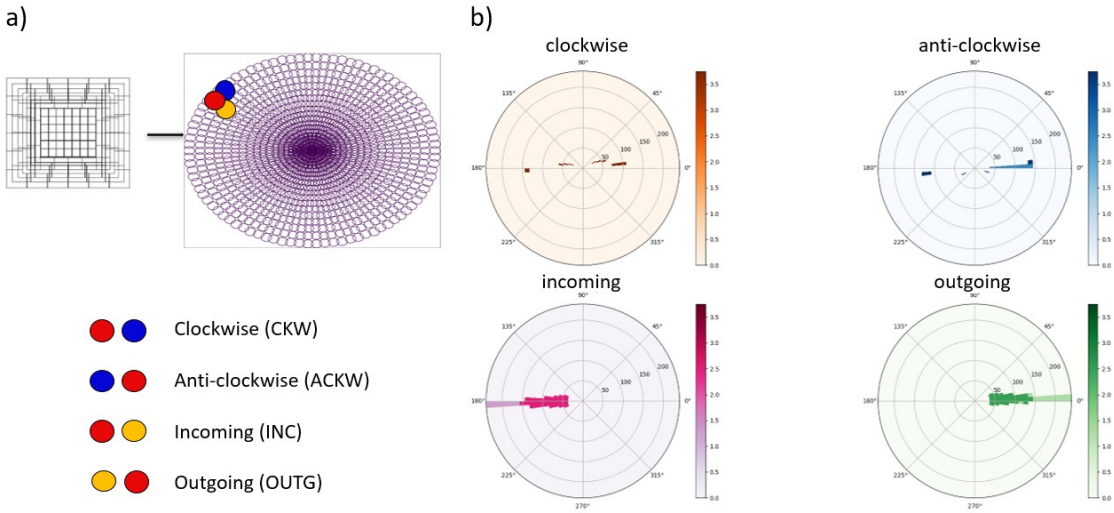


Figure 5.13. **a)** Schematic representation of the log-polar architecture. a) Difference between the ”eccentric” subsample proposed in Ch. 5 and the ongoing log-polar implementation. The four cardinal direction changed to: incoming and outgoing the fovea, clockwise and anti-clockwise. b) Log-polar representation of the log-polar model response (MFR) to a bar moving from left to right at the center of the visual field.

Chapter 6

Discussion

The proto-object saliency-based model demonstrated to be a suitable biological plausible system to detect salient regions of the scene allowing extension for different channels and sources of information to be integrated, playing a role in the generation of the saliency map. The saliency-based proto-object model is based on the Gestalt intuition to introduce the Border Ownership concept in visual perception. The model effectively detects the presence of a possible object in the scene discarding the clutter in static and dynamic scenes. The event-driven system drastically reduced the need for computation due to the inherent properties of the event-driven sensors. The use of these cameras leveraged the system's capabilities to produce an outcome with a reduction in latency and computational load. Whilst the event-based proto-object model does not completely get rid of the clutter, the spiking-based intensity channel effectively removes the clutter due to ON and OFF polarity spikes balancing each other avoiding triggering the Von Mises correspondent neuron. These results prove the advantage of fully spike-based pipelines, enhanced by the significant reduction in latency, to obtain output spikes from the system. In both cases, event-driven and fully spiking-based implementation, the intensity channel does not seem to robustly focus on a target, as it jumps from one proto-object to another one depending on the number of events. This behaviour can be corrected by adding a Winner-Take-All mechanism with hysteresis avoiding the same focus multiple times in a short period of time. That gives a competitive advantage to the current winner and stability to the selection. The attentional scan path is then generated by the complex interplay between WTA and inhibition-of-return. The response of both models is dependent on the number of events and indirectly on the intrinsic information of motion and size of the entities in the scene. For the intensity channel, these factors should be modulated giving equal priority to objects with different sizes and motions unless the task is clearly defined. This simple problem gives rise to an important and focal question around attention.
Do bottom-up and top-down mechanisms follow separate paths in the entire attention process or do they contribute and compete with each other at the same time towards the representation of the scene? The interesting question behind this is, why do we need to split

these two pathways forcing the saliency-based models to be purely bottom-up or top-down? An answer to this issue lies in the complexity of these mechanisms to be controlled allowing a natural response from a global schema. These two processes are indeed not separated [141], making the creation of an attention schema a critical challenge. The interplay between data-driven and task-driven visual attention mechanisms represents a complex open topic. This scientific question requires a in detail analysis looking at human mechanisms. Indeed, humans can be alert while performing top-down visual search tasks seemingly without effort. For the depth channel, I assumed a precise task for the robot to focus attention towards entities in its closeness. This approach allowed me to identify a specific task by prioritising objects with which the robot can interact because of their proximity. We proved depth to be an important cue for the proto-object model. The evProtoDepth model proved the need for the combination of depth perception with proto-object detection. Depth perception alone cannot guarantee the perception of actual objects in the scene. The presence of the Gestalt-based proto-object model is, therefore, necessary to allow the interaction of the robot with close items. The online system detects proto-objects without being driven by the number of events prioritising proto-objects over random clusters with high disparity. This is by far the most relevant result regarding this channel due to the choice to directly feed the disparity map into the Border Ownership Pyramid. The model suits the addition of different channels such as texture [89] and orientation [9]. These channels could be merged, prioritising and modulating their influence depending on the task. The proposed models are feedforward architectures that could be enriched with feedback connections to strengthen proto-object detection. A future channel could be added to represent the figure-ground segmentation [181]. The event-driven figure-ground organisation model has already been implemented in Python, showing promising initial results (see Figure 6.1). This model distinguishes the foreground from the background by exploiting feedforward and feedback connections.

One of the main limitations of the proposed saliency-based visual attention model is the impossibility to obtain a ground truth comparable with the purely bottom-up system we proposed. Human fixation maps are the result of combined top-down and bottom-up mechanisms making the available benchmark datasets good only to validate the work, assessing how far the response from the model is from the actual ground truth. The problem depends on the several saliency-based metrics existent in the literature complicating the validation of the system. The fixation maps used as ground truth come from more complex mechanisms than the mechanisms represented in the proposed bottom-up model. This fundamental difference does not guarantee a fair comparison even if the resulting saliency maps would be equal. The experiments are a model validation verifying the distance from the real ground truth. The fixation maps from human subjects are the only source of ground

truth. Therefore, a single saliency map cannot perform well in all the metrics, especially because each metric is looking for a specific aspect of the saliency map. These metrics rely on different factors of fixation points treating differently false positives and negatives. This analysis causes a saliency map that is optimal for one metric to perform worse than the baselines in other metrics [140].

The motion channel is the only channel where the output does not input into the Border Ownership Pyramid. This model proved the real applicability of biologically plausible pipelines by taking advantage of a retina-like structure where the receptive fields have different sizes across the visual field. This model detects the direction of motion with a lower firing rate from the neurons, despite the increased number of neurons required for robustness. This work presents two main results, the wider speed range obtained making the retina layer able to detect different speeds depending on the areas of the visual field and the possibility to detect transversal directions combining the response from the neuron populations. Fast speeds are detected from the receptive fields in the periphery of the visual field allowing a fast reaction of the robot to incoming stimuli, effectively making the robot able to be in an alert state. The motion information does not input into the proto-object model. In fact, motion alone can already independently trigger visual alertness in attention mechanisms. Motion is per se an attentional cue [52]. Furthermore, the onset of coherent motion attracts attention regardless of the luminance change [52] suggesting a different processing from the Border Ownership cells. Indeed, motion can be detected without attention and it is considered a fundamental component in early vision [241]. This motion detector model for the spiking-based version of the proto-object implementation, both on SpiNNaker, are affected by the same problem. These platforms allow low latency and reduced power consumption due to the spike-based architecture but they show limitations in scalability. One of the challenges faced by neuromorphic computing is the limitation on the possible number of neurons and synapses achievable by the currently available platforms. The SpiNNaker platform (SpiNN-5, 48 chips) can simulate up to 80.000 neurons and 0.3 billion synapses [242]. BrainScale-2 provides an analog core with 512 neurons and 2^7 synapses [243]. Loihi offers 128 cores each containing 1024 neural primitives units [130]. DyNAPS supports event-based neural networks providing 1k as the maximum number of neurons and 64k number of synapses [131]. All of these platforms provide a significant amount of fixed possible neurons enabling the creation of complex networks. The full resolution of the ATIS cameras used for this project is 304x240 pixels, requiring a high number of neurons in input (72,960). The possible applications exploiting neuromorphic hardware need to down-sample the visual field or make use of small networks avoiding the creation of multiple populations. The analog platforms (DyNAPS, BrainScale-2) do not offer a reasonable amount of neurons to allow an acceptable size of the visual field.

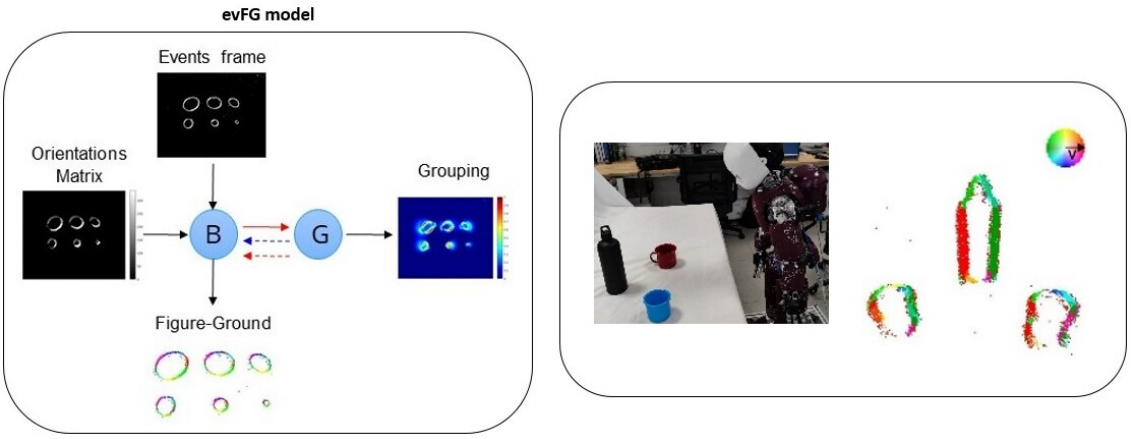


Figure 6.1. On the left: a scheme of the Event-Driven Figure-Ground organisation model (evFG) taking inspiration from the RGB implementation [181]. On the right: the first experiments distinguishing the foreground. The model detects objects pointing an arrow toward the object centre (see the Legend at the top-right).

SpiNNaker and Loihi represent a feasible choice to implement visual algorithms. While SpiNNaker works building connections among neurons where each neuron (soma) can fire emitting spikes, in Loihi, any compartment can fire. Each element of the network (axon, dendrite, and soma) is represented as a compartment allowing great flexibility. Loihi and SpiNNaker are fully clock-based digital both accepting spike arrays as input. Loihi is working with a defined timestep period and SpiNNaker exploits timestamps. Both SpiNNaker and Loihi allow modelling neurons granting flexibility. These platforms indeed support the exploration of bioinspired architectures, they both would need a more user-friendly experience providing for example a set of matrix connections already prepared or either the possibility to draw the network. Neuromorphic computing represents a valid and worth exploring alternative to classical computing. The inherent capabilities of event-driven cameras, along with the cut out of redundant information, allowed the remotion of processing layers. The low latency obtained, due to the SpiNNaker board shown from the SNNevProto and the Eccentric sEMD clearly confirms the benefit of the neuromorphic hardware. Robotic applications seeking fast responses in unconstrained environments can thoroughly take advantage of this new generation of computing allowing for a ”natural” response from the robot.

Chapter 7

Conclusions

This work answers the initial scientific question arose at the start of the project.

The bridge between bioinspired hardware and software is possible, and it helps the reduction in computational loads and latency needed to obtain a response from the system.

The pipelines which take advantage of the fully spike-based approach, the SNN Intensity channel and the Motion channel, reduced dramatically the latency, at the cost of the complexity of the model. The number of SpiNNaker boards needed depending on the number of neurons required is still not convenient for an application where the robot needs to move around.
The Intensity and the Depth channel run online on the robot exploiting C++ and PyTorch on a GPU, providing the system with an outcome in 100/200ms, approximately the same time needed to perform a saccade [7], [8]. This is the first big achievement compared to the RGB model I took inspiration from running on Matlab, requiring minutes for the response. Thanks to this latency the robot can realistically interact with external stimuli.
The model exploits a fully bio-inspired biologically plausible pipeline making the basis for a complex scenario where the robot needs to shift attention depending on different tasks yet be able to remain alert. This work aims to start a complex attention schema where intensity, depth and motion are the initial channels where several other sources of information can be added. The start of the complex attention schema would be to weigh and prioritise one channel at a time, or a combination of channels, to obtain a specific response for a particular task. The modulation of contributions among different channels is still an open question due to the combination of bottom-up and top-down mechanisms. This project certainly lacks a WTA mechanism to select the maximum value representing the salient location [73] based on the Intensity, Depth and Motion channel. Once the model finally finds a point to fixate on, the inhibition-of-return mechanism becomes necessary to allow the next shift towards the new salient location. The limitations of attention-based models are the impossibility to have a ground truth to assess the system or define a loss function for training. Human fixations represent the only ground truth available to evaluate

a saliency map response. The cognitive bias behind a human attention focus cannot be clearly explained from a purely bottom-up or top-down perspective [141]. The deep interplay between these two mechanisms does not guarantee voluntary human fixations over a specific point not even if the task is chosen. Obtaining a loss function would allow learning the kernel used for proto-object detection or learning the saliency map for a specific attention task. Learning the kernel would discover new kernels for specific tasks or confirm the used curved filter. Training on human fixation maps to move the robot's eyes would learn fixation points depending on the training data. This would not solve the question about the bottom-up mechanisms behind the attention dilemma. The event-based approach is fast, efficient and sparse, unlocking new possibilities for attention models in robotic applications. Future perspectives see many challenges in the combination of bottom-up and top-down mechanisms working together towards an autonomous selection of the salient point. How to switch from one mechanism to another one? Also, do they have to work separately? How much a bottom-up attentional cue interferes when the model is performing a task?

All of these questions can be explored by different experts analysing human attentional reactions to different situations. To answer such scientific questions a fundamental step would also be the creation of the first event-driven saliency-based dataset providing the ground truth for event-based approaches.

References

[1] G. Metta, G. Sandini, D. Vernon, L. Natale, and F. Nori, "The icub humanoid robot: An open platform for research in embodied cognition," in *Proceedings of the 8th workshop on performance metrics for intelligent systems*, 2008, pp. 50–56.

[2] C. Bartolozzi, F. Rea, C. Clercq, D. B. Fasnacht, G. Indiveri, M. Hofstätter, and G. Metta, "Embedded neuromorphic vision for humanoid robots," in *CVPR 2011 workshops*, IEEE, 2011, pp. 129–135.

[3] F. Briggs, "Mammalian visual system organization," in *Oxford Research Encyclopedia of Neuroscience*, 2017.

[4] J. M. Wolfe, "Visual attention," *Seeing*, pp. 335–386, 2000.

[5] C. Posch, D. Matolin, and R. Wohlgenannt, "A QVGA 143 dB dynamic range frame-free PWM image sensor with lossless pixel-level video compression and time-domain CDS," in *IEEE Journal of Solid-State Circuits*, vol. 46, Jan. 2011, pp. 259–275, ISBN: 9781424460342. DOI: 10.1109/JSSC.2010.2085952. [Online]. Available: http://ieeexplore.ieee.org/document/5648367/.

[6] S. B. Furber, F. Galluppi, S. Temple, and L. A. Plana, "The spinnaker project," *Proceedings of the IEEE*, vol. 102, no. 5, pp. 652–665, May 2014, ISSN: 0018-9219. DOI: 10.1109/JPROC.2014.2304638.

[7] B. Fischer and R. Boch, "Saccadic eye movements after extremely short reaction times in the monkey," *Brain research*, vol. 260, no. 1, pp. 21–26, 1983.

[8] B. Fischer, M. Biscaldi, and S. Gezeck, "On the development of voluntary and reflexive components in human saccade generation," *Brain research*, vol. 754, no. 1-2, pp. 285–297, 1997.

[9] A. F. Russell, S. Mihalaş, R. von der Heydt, E. Niebur, and R. Etienne-Cummings, "A model of proto-object based saliency," *Vision research*, vol. 94, pp. 1–15, 2014.

[10] M. B. Milde, O. J. Bertrand, H. Ramachandran, M. Egelhaaf, and E. Chicca, "Spiking elementary motion detector in neuromorphic systems," *Neural computation*, vol. 30, no. 9, pp. 2384–2417, 2018.

[11] M. Firouzi and J. Conradt, "Asynchronous event-based cooperative stereo matching using neuromorphic silicon retinas," *Neural Processing Letters*, vol. 43, no. 2, pp. 311–326, 2016.

[12] A. Glover, A. B. Stokes, S. Furber, and C. Bartolozzi, "Atis+ spinnaker: A fully event-based visual tracking demonstration," *arXiv preprint arXiv:1912.01320*, 2019.

[13] M. B. Milde, H. Blum, A. Dietmüller, D. Sumislawska, J. Conradt, G. Indiveri, and Y. Sandamirskaya, "Obstacle avoidance and target acquisition for robot navigation using a mixed signal analog/digital neuromorphic processing system," *Frontiers in neurorobotics*, vol. 11, p. 28, 2017.

[14] M. Thor, B. Strohmer, and P. Manoonpong, "Locomotion control with frequency and motor pattern adaptations," *Frontiers in Neural Circuits*, vol. 15, 2021.

[15] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, *et al.*, "Event-based vision: A survey," *IEEE transactions on pattern analysis and machine intelligence*, vol. 44, no. 1, pp. 154–180, 2020.

[16] G. Indiveri, F. Corradi, and N. Qiao, "Neuromorphic architectures for spiking deep neural networks," in *2015 IEEE International Electron Devices Meeting (IEDM)*, IEEE, 2015, pp. 4–2.

[17] P. A. Bogdan, B. Marcinnò, C. Casellato, S. Casali, A. G. Rowley, M. Hopkins, F. Leporati, E. D'Angelo, and O. Rhodes, "Towards a bio-inspired real-time neuromorphic cerebellum," *Frontiers in cellular neuroscience*, vol. 15, p. 622 870, 2021.

[18] G. Orchard, X. Lagorce, C. Posch, S. B. Furber, R. Benosman, and F. Galluppi, "Real-time event-driven spiking neural network object recognition on the spinnaker platform," in *2015 IEEE International Symposium on Circuits and Systems (ISCAS)*, IEEE, 2015, pp. 2413–2416.

[19] A. M. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive psychology*, vol. 12, no. 1, pp. 97–136, 1980.

[20] W. James, F. Burkhardt, F. Bowers, and I. K. Skrupskelis, *The principles of psychology*, 2. Macmillan London, 1890, vol. 1.

[21] G. Rizzolatti, "Mechanisms of selective attention in mammals," in *Advances in vertebrate neuroethology*, Springer, 1983, pp. 261–297.

[22] H. Pashler, *Attention*. Psychology Press, 2016.

[23] B. W. Tatler, N. J. Wade, H. Kwan, J. M. Findlay, and B. M. Velichkovsky, "Yarbus, eye movements, and vision," *i-Perception*, vol. 1, no. 1, pp. 7–27, 2010. [Online]. Available: `https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3563050/`.

[24] A. L. Yarbus, "Eye movements during perception of complex objects," in *Eye movements and vision*, Springer, 1967, pp. 171–211.

[25] ——, *Eye movements and vision*. Springer, 2013.

[26] B. J. Scholl, "Objects and attention: The state of the art," *Cognition*, vol. 80, no. 1-2, pp. 1–46, 2001.

[27] H. Von Helmholtz, "Treatise on physiological optics vol. iii," 1867.

[28] M. I. Posner, "Orienting of attention," *Quarterly journal of experimental psychology*, vol. 32, no. 1, pp. 3–25, 1980.

[29] H. E. Egeth and S. Yantis, "Visual attention: Control, representation, and time course," *Annual review of psychology*, vol. 48, no. 1, pp. 269–297, 1997.

[30] R. M. McPeek and K. Nakayama, "Linkage of attention and saccades in a visual search task," *Investigative Ophthalmology and Visual Science*, vol. 36, no. 4, S354, 1995.

[31] M. P. Bryden, "The role of post-exposural eye movements in tachistoscopic perception.," *Canadian Journal of Psychology/Revue canadienne de psychologie*, vol. 15, no. 4, p. 220, 1961.

[32] H. F. Crovitz and W. Daves, "Tendencies to eye movement and perceptual accuracy.," *Journal of Experimental Psychology*, vol. 63, no. 5, p. 495, 1962.

[33] 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*, vol. 20, no. 11, pp. 1254–1259, 1998.

[34] D. J. Parkhurst, *Selective attention in natural vision: Using computational models to quantify stimulus-driven attentional allocation*. The Johns Hopkins University, 2002.

[35] A. Borji, D. N. Sihite, and L. Itti, "Quantitative analysis of human-model agreement in visual saliency modeling: A comparative study," *Image Processing, IEEE Transactions on*, vol. 22, no. 1, pp. 55–69, 2013.

[36] M. I. Posner and Y. Cohen, "Attention and the control of movements," in *Advances in Psychology*, vol. 1, Elsevier, 1980, pp. 243–258.

[37] C. E. Connor, H. E. Egeth, and S. Yantis, "Visual attention: Bottom-up versus top-down," *Current biology*, vol. 14, no. 19, R850–R852, 2004.

[38] K. Kompatsiari, F. Ciardo, D. De Tommaso, and A. Wykowska, "Measuring engagement elicited by eye contact in human-robot interaction," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 6979–6985. DOI: 10.1109/IROS40897.2019.8967747.

[39] K. Kompatsiari, J. Pérez-Osorio, D. De Tommaso, G. Metta, and A. Wykowska, "Neuroscientifically-grounded research for improved human-robot interaction," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2018, pp. 3403–3408. DOI: 10.1109/IROS.2018.8594441.

[40] C. Breazeal and B. Scassellati, "A context-dependent attention system for a social robot," *rn*, vol. 255, no. 3, 1999.

[41] A. Ude, V. Wyart, L.-H. Lin, and G. Cheng, "Distributed visual attention on a humanoid robot," in *5th IEEE-RAS International Conference on Humanoid Robots, 2005.*, IEEE, 2005, pp. 381–386.

[42] S. Vijayakumar, J. Conradt, T. Shibata, and S. Schaal, "Overt visual attention for a humanoid robot," in *Proceedings 2001 IEEE/RSJ International Conference on Intelligent Robots and Systems. Expanding the Societal Role of Robotics in the the Next Millennium (Cat. No. 01CH37180)*, IEEE, vol. 4, 2001, pp. 2332–2337.

[43] J. Ruesch, M. Lopes, A. Bernardino, J. Hornstein, J. Santos-Victor, and R. Pfeifer, "Multimodal saliency-based bottom-up attention a framework for the humanoid robot icub," in *2008 IEEE International Conference on Robotics and Automation*, IEEE, 2008, pp. 962–967.

[44] F. Rea, G. Metta, and C. Bartolozzi, "Event-driven visual attention for the humanoid robot icub," *Frontiers in neuroscience*, vol. 7, p. 234, 2013.

[45] N. Kawabata, "Attention and depth perception," *Perception*, vol. 15, no. 5, pp. 563–572, 1986.

[46] J. J. Clark and N. J. Ferrier, "Modal control of an attentive vision system.," in *ICCV*, 1988, pp. 514–523.

[47] K. Pahlavan, T. Uhlin, and J.-O. Eklundh, "Integrating primary ocular processes," in *European Conference on Computer Vision*, Springer, 1992, pp. 526–541.

[48] N. D. Bruce and J. K. Tsotsos, "An attentional framework for stereo vision," in *The 2nd Canadian Conference on Computer and Robot Vision (CRV'05)*, IEEE, 2005, pp. 88–95.

[49] G. Pasquale, T. Mar, C. Ciliberto, L. Rosasco, and L. Natale, "Enabling depth-driven visual attention on the icub humanoid robot: Instructions for use and new perspectives," *Frontiers in Robotics and AI*, vol. 3, p. 35, 2016.

[50] J. Thompson and R. Parasuraman, "Attention, biological motion, and action recognition," *Neuroimage*, vol. 59, no. 1, pp. 4–13, 2012.

[51] P. Cavanagh, "Attention-based motion perception," *Science*, vol. 257, no. 5076, pp. 1563–1565, 1992.

[52] R. A. Abrams and S. E. Christ, "Motion onset captures attention," *Psychological Science*, vol. 14, no. 5, pp. 427–432, 2003.

[53] Y.-F. Ma and H.-J. Zhang, "A model of motion attention for video skimming," in *Proceedings. International Conference on Image Processing*, vol. 1, 2002, pp. I–I. DOI: 10.1109/ICIP.2002.1037976.

[54] S. Chen and Y.-G. Jiang, "Motion guided spatial attention for video captioning," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 33, 2019, pp. 8191–8198.

[55] H. Li, G. Chen, G. Li, and Y. Yu, "Motion guided attention for video salient object detection," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7274–7283.

[56] T. Fraichard, R. Paulin, and P. Reignier, "Human-robot motion: An attention-based navigation approach," in *The 23rd IEEE International Symposium on Robot and Human Interactive Communication*, 2014, pp. 684–691. DOI: 10.1109/ROMAN.2014.6926332.

[57] Q. Zhao and C. Koch, "Learning saliency-based visual attention: A review," *Signal Processing*, vol. 93, no. 6, pp. 1401–1407, 2013.

[58] S. Minut and S. Mahadevan, "A reinforcement learning model of selective visual attention," 2001.

[59] U. Rutishauser, D. Walther, C. Koch, and P. Perona, "Is bottom-up attention useful for object recognition?" In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004. CVPR 2004.*, IEEE, vol. 2, 2004, pp. II–II.

[60] V. Mnih, N. Heess, A. Graves, *et al.*, "Recurrent models of visual attention," *Advances in neural information processing systems*, vol. 27, 2014.

[61] N. Tong, H. Lu, X. Ruan, and M.-H. Yang, "Salient object detection via bootstrap learning," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, IEEE, 2015, pp. 1884–1892.

[62] P.-H. Tseng, R. Carmi, I. G. Cameron, D. P. Munoz, and L. Itti, "Quantifying center bias of observers in free viewing of dynamic natural scenes," *Journal of vision*, vol. 9, no. 7, pp. 4–4, 2009.

[63] M. Kummerer, T. S. Wallis, L. A. Gatys, and M. Bethge, "Understanding low- and high-level contributions to fixation prediction," in *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 4789–4798.

[64] B. Guo, N. Guo, and Z. Cen, "Motion saliency-based collision avoidance for mobile robots in dynamic environments," *IEEE Transactions on Industrial Electronics*, 2021.

[65] B. Schauerte, B. Kühn, K. Kroschel, and R. Stiefelhagen, "Multimodal saliency-based attention for object-based scene analysis," in *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2011, pp. 1173–1179.

[66] N. J. Butko, L. Zhang, G. W. Cottrell, and J. R. Movellan, “Visual saliency model for robot cameras,” in *2008 IEEE International Conference on Robotics and Automation*, IEEE, 2008, pp. 2398–2403.

[67] Z. Ma, C. Wang, Y. Niu, X. Wang, and L. Shen, “A saliency-based reinforcement learning approach for a uav to avoid flying obstacles,” *Robotics and Autonomous Systems*, vol. 100, pp. 108–118, 2018.

[68] S.-J. Park, J.-K. Shin, and M. Lee, “Biologically inspired saliency map model for bottom-up visual attention,” in *International Workshop on Biologically Motivated Computer Vision*, Springer, 2002, pp. 418–426.

[69] S. He, J. Han, X. Hu, M. Xu, L. Guo, and T. Liu, “A biologically inspired computational model for image saliency detection,” in *Proceedings of the 19th ACM international conference on Multimedia*, 2011, pp. 1465–1468.

[70] V. Mahadevan and N. Vasconcelos, “Biologically inspired object tracking using center-surround saliency mechanisms,” *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 3, pp. 541–554, 2012.

[71] J. Zhao, S. Sun, X. Liu, J. Sun, and A. Yang, “A novel biologically inspired visual saliency model,” *Cognitive Computation*, vol. 6, no. 4, pp. 841–848, 2014.

[72] L. Duan, J. Gu, Z. Yang, J. Miao, W. Ma, and C. Wu, “Bio-inspired visual attention model and saliency guided object segmentation,” in *Genetic and Evolutionary Computing*, Springer, 2014, pp. 291–298.

[73] C. Bartolozzi and G. Indiveri, “Selective attention implemented with dynamic synapses and integrate-and-fire neurons,” *Neurocomputing*, vol. 69, no. 16-18, pp. 1971–1976, 2006.

[74] ——, “Selective attention in multi-chip address-event systems,” *Sensors*, vol. 9, no. 7, pp. 5076–8098, 2009.

[75] L. Itti and C. Koch, “Computational modelling of visual attention,” *Nature reviews neuroscience*, vol. 2, no. 3, pp. 194–203, 2001.

[76] K. Koffka and Á. Cabral, *Princípios de psicologia da Gestalt*. Cultrix São Paulo, 1975.

[77] J. Wagemans, J. Feldman, S. Gepshtein, R. Kimchi, J. R. Pomerantz, P. A. Van der Helm, and C. Van Leeuwen, "A century of gestalt psychology in visual perception: Ii. conceptual and theoretical foundations.," *Psychological bulletin*, vol. 138, no. 6, p. 1218, 2012.

[78] P. A. Kolers, *Aspects of motion perception: International series of monographs in experimental psychology*. Elsevier, 2013, vol. 16.

[79] J. Hochberg, *Perception and Cognition at Century's End: History, Philosophy, Theory*. Elsevier, 1998.

[80] E. Hering, *Beitrage zur physiologie*. W. Engelmann, 1861.

[81] E. Mach, *The analysis of sensations, and the relation of the physical to the psychical*. Open Court Publishing Company, 1914.

[82] C. v. Ehrenfels, "Über gestaltqualitäten," *Vierteljahrsschrift für wissenschaftliche Philosophie*, vol. 14, no. 3, pp. 249–292, 1890.

[83] W. Köhler, "Gestalt psychology," *Psychological research*, vol. 31, no. 1, pp. XVIII–XXX, 1967.

[84] P. W. Besslich and H. Bässmann, "Gestalt-based approach to robot vision," *Expert systems and robotics*, pp. 1–34, 1991.

[85] F. Orabona, G. Metta, and G. Sandini, "A proto-object based visual attention model," in *International Workshop on Attention in Cognitive Systems*, Springer, 2007, pp. 198–215.

[86] M. Pardowitz, R. Haschke, J. Steil, and H. Ritter, "Gestalt-based action segmentation for robot task learning," in *Humanoids 2008-8th IEEE-RAS International Conference on Humanoid Robots*, IEEE, 2008, pp. 347–352.

[87] J. L. Molin, A. F. Russell, S. Mihalas, E. Niebur, and R. Etienne-Cummings, "Proto-object based visual saliency model with a motion-sensitive channel," in *2013 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2013, pp. 25–28.

[88] B. Hu, R. Kane-Jackson, and E. Niebur, "A proto-object based saliency model in three-dimensional space," *Vision research*, vol. 119, pp. 42–49, 2016.

[89] T. Uejima, E. Niebur, and R. Etienne-Cummings, "Proto-object based saliency model with second-order texture feature," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–4.

[90] B. Hu and E. Niebur, "A recurrent neural model for proto-object based contour integration and figure-ground segregation," *Journal of computational neuroscience*, vol. 43, no. 3, pp. 227–242, 2017.

[91] J. Schooler, "Bridging the objective/subjective divide: Towards a meta-perspective of science and experience," in *Open MIND*, Open MIND. Frankfurt am Main: MIND Group, 2014.

[92] H. Zhou, H. S. Friedman, and R. Von Der Heydt, "Coding of border ownership in monkey visual cortex," *Journal of Neuroscience*, vol. 20, no. 17, pp. 6594–6611, 2000.

[93] E. Rubin, "Figure and ground," 2001.

[94] D. Hoiem, A. A. Efros, and M. Hebert, "Recovering occlusion boundaries from an image," *International Journal of Computer Vision*, vol. 91, no. 3, pp. 328–346, 2011.

[95] E. Borenstein and S. Ullman, "Combined top-down/bottom-up segmentation," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 30, no. 12, pp. 2109–2125, 2008.

[96] R. von der Heydt, T. Macuda, and F. T. Qiu, "Border-ownership-dependent tilt aftereffect," *JOSA A*, vol. 22, no. 10, pp. 2222–2229, 2005.

[97] F. Fang, H. Boyaci, and D. Kersten, "Border ownership selectivity in human early visual cortex and its modulation by attention," *Journal of Neuroscience*, vol. 29, no. 2, pp. 460–465, 2009.

[98] J. R. Williford and R. von der Heydt, "Border-ownership coding," *Scholarpedia journal*, vol. 8, no. 10, p. 30 040, 2013.

[99] P. Mehrani and J. K. Tsotsos, "Early recurrence enables figure border ownership," *Vision Research*, vol. 186, pp. 23–33, 2021.

[100] R. A. Rensink, "Seeing, sensing, and scrutinizing," *Vision research*, vol. 40, no. 10-12, pp. 1469–1487, 2000.

[101] ——, "The dynamic representation of scenes," *Visual cognition*, vol. 7, no. 1-3, pp. 17–42, 2000.

[102] D. Walther and C. Koch, "Modeling attention to salient proto-objects," *Neural networks*, vol. 19, no. 9, pp. 1395–1407, 2006.

[103] M. Riesenhuber and T. Poggio, "Hierarchical models of object recognition in cortex," *Nature neuroscience*, vol. 2, no. 11, pp. 1019–1025, 1999.

[104] V. Yanulevskaya, J. Uijlings, J.-M. Geusebroek, N. Sebe, and A. Smeulders, "A proto-object-based computational model for visual saliency," *Journal of vision*, vol. 13, no. 13, pp. 27–27, 2013.

[105] C.-P. Yu, D. Samaras, and G. J. Zelinsky, "Modeling visual clutter perception using proto-object segmentation," *Journal of vision*, vol. 14, no. 7, pp. 4–4, 2014.

[106] E. H. Adelson and J. R. Bergen, "Spatiotemporal energy models for the perception of motion," *Josa a*, vol. 2, no. 2, pp. 284–299, 1985.

[107] M. Concetta Morrone and D. Burr, "Feature detection in human vision: A phase-dependent energy model," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 235, no. 1280, pp. 221–245, 1988.

[108] E. H. Adelson, C. H. Anderson, J. R. Bergen, P. J. Burt, and J. M. Ogden, "Pyramid methods in image processing," *RCA engineer*, vol. 29, no. 6, pp. 33–41, 1984.

[109] E. Mancinelli, E. Niebur, and R. Etienne-Cummings, "Computational stereo-vision model of proto-object based saliency in three-dimensional space," in *2018 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, IEEE, 2018, pp. 1–4.

[110] E. Mueggler, B. Huber, and D. Scaramuzza, "Event-based, 6-dof pose tracking for high-speed maneuvers," in *2014 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2014, pp. 2761–2768.

[111] V. I. Govardovskii, P. D. Calvert, and V. Y. Arshavsky, "Photoreceptor light adaptation: Untangling desensitization and sensitization," *The Journal of General Physiology*, vol. 116, no. 6, pp. 791–794, 2000.

[112] B. Scholl, K. W. Latimer, and N. J. Priebe, "A retinal source of spatial contrast gain control," *Journal of Neuroscience*, vol. 32, no. 29, pp. 9824–9830, 2012.

[113] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128x128 120 db 15us latency asynchronous temporal contrast vision sensor," *IEEE journal of solid-state circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[114] A. Glover and C. Bartolozzi, "Event-driven ball detection and gaze fixation in clutter," in *IEEE International Conference on Intelligent Robots and Systems*, vol. 2016-Novem, IEEE, 2016, pp. 2203–2208, ISBN: 9781509037629. DOI: 10.1109/IROS.2016.7759345. [Online]. Available: http://ieeexplore.ieee.org/document/7759345/.

[115] ——, "Robust visual tracking with a freely-moving event camera," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2017, pp. 3769–3776.

[116] M. Monforte, A. Arriandiaga, A. Glover, and C. Bartolozzi, "Exploiting event cameras for spatio-temporal prediction of fast-changing trajectories," in *2020 2nd IEEE International Conference on Artificial Intelligence Circuits and Systems (AICAS)*, IEEE, 2020, pp. 108–112.

[117] V. Vasco, A. Glover, and C. Bartolozzi, "Fast event-based harris corner detection exploiting the advantages of event-driven cameras," in *2016 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2016, pp. 4144–4149.

[118] M. Iacono, S. Weber, A. Glover, and C. Bartolozzi, "Towards Event-Driven Object Detection with Off-The-Shelf Deep Learning," in *IEEE International Conference on Intelligent Robots and Systems*, IROS 2018, 2018.

[119] T. Stoffregen, G. Gallego, T. Drummond, L. Kleeman, and D. Scaramuzza, "Event-based motion segmentation by motion compensation," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 7244–7253.

[120] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3867–3876.

[121] G. Gallego and D. Scaramuzza, "Accurate angular velocity estimation with an event camera," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 632–639, 2017.

[122] F. Barranco, C. Fermüller, and Y. Aloimonos, "Contour motion estimation for asynchronous event-driven cameras," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1537–1556, 2014.

[123] T. Schoepe, E. Janotte, M. B. Milde, O. J. Bertrand, M. Egelhaaf, and E. Chicca, "Finding the gap: Neuromorphic motion vision in cluttered environments," *arXiv preprint arXiv:2102.08417*, 2021.

[124] F. Barranco, C. Fermuller, and Y. Aloimonos, "Bio-inspired motion estimation with event-driven sensors," in *International Work-Conference on Artificial Neural Networks*, Springer, 2015, pp. 309–321.

[125] M. Osswald, S.-H. Ieng, R. Benosman, and G. Indiveri, "A spiking neural network model of 3d perception for event-based neuromorphic stereo vision systems," *Scientific reports*, vol. 7, no. 1, pp. 1–12, 2017.

[126] D. Monroe, *Neuromorphic computing gets ready for the (really) big time*, 2014.

[127] A. L. Hodgkin and A. F. Huxley, "A quantitative description of membrane current and its application to conduction and excitation in nerve," *The Journal of physiology*, vol. 117, no. 4, p. 500, 1952.

[128] A. LF, "Lapicque's introduction of the integrate-and-fire model neuron (1907," *Brain Res Bull*, no. 50, pp. 5–6, 1999.

[129] M. Mahowald and R. Douglas, "A silicon neuron," *Nature*, vol. 354, no. 6354, pp. 515–518, 1991.

[130] M. Davies, N. Srinivasa, T.-H. Lin, G. Chinya, Y. Cao, S. H. Choday, G. Dimou, P. Joshi, N. Imam, S. Jain, *et al.*, "Loihi: A neuromorphic manycore processor with on-chip learning," *Ieee Micro*, vol. 38, no. 1, pp. 82–99, 2018.

[131] S. Moradi, N. Qiao, F. Stefanini, and G. Indiveri, "A scalable multicore architecture with heterogeneous memory structures for dynamic neuromorphic asynchronous processors (dynaps)," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 1, pp. 106–122, 2017.

[132] R. Massa, A. Marchisio, M. Martina, and M. Shafique, "An efficient spiking neural network for recognizing gestures with a dvs camera on the loihi neuromorphic processor," in *2020 International Joint Conference on Neural Networks (IJCNN)*, IEEE, 2020, pp. 1–9.

[133] N. Risi, A. Aimar, E. Donati, S. Solinas, and G. Indiveri, "A spike-based neuromorphic architecture of stereo vision," *Frontiers in neurorobotics*, vol. 14, p. 568 283, 2020.

[134] Q. Liu, D. Xing, L. Feng, H. Tang, and G. Pan, "Event-based multimodal spiking neural network with attention mechanism," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2022, pp. 8922–8926.

[135] A. Gruel and J. Martinet, "Bio-inspired visual attention for silicon retinas based on spiking neural networks applied to pattern classification," in *2021 International Conference on Content-Based Multimedia Indexing (CBMI)*, IEEE, 2021, pp. 1–6.

[136] M. Yao, H. Gao, G. Zhao, D. Wang, Y. Lin, Z. Yang, and G. Li, "Temporal-wise attention spiking neural networks for event streams classification," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 10 221–10 230.

[137] Z. Bylinskii, T. Judd, A. Oliva, A. Torralba, and F. Durand, "What do different evaluation metrics tell us about saliency models?" *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 41, no. 3, pp. 740–757, 2019. DOI: 10.1109/TPAMI.2018.2815601.

[138] T. Judd, F. Durand, and A. Torralba, "A benchmark of computational models of saliency to predict human fixations," in *MIT Technical Report*, 2012.

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

[140] M. Kummerer, T. S. Wallis, and M. Bethge, "Saliency benchmarking made easy: Separating models, maps and metrics," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 770–787.

[141] A. Wykowska and A. Schubö, "On the temporal relation of top–down and bottom–up mechanisms during guidance of attention," *Journal of Cognitive Neuroscience*, vol. 22, no. 4, pp. 640–654, 2010.

[142] B. Hassenstein and W. Reichardt, "Systemtheoretische analyse der zeit-, reihenfolgen- und vorzeichenauswertung bei der bewegungsperzeption des rüsselkäfers chlorophanus," *Zeitschrift für Naturforschung B*, vol. 11, no. 9-10, pp. 513–524, 1956.

[143] H. Barlow and W. R. Levick, "The mechanism of directionally selective units in rabbit's retina.," *The Journal of physiology*, vol. 178, no. 3, pp. 477–504, 1965.

[144] J. Kramer, "Compact integrated motion sensor with three-pixel interaction," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 18, no. 4, pp. 455–460, 1996.

[145] S. P. Strong, R. Koberle, R. R. de Ruyter van Steveninck, and W. Bialek, "Entropy and Information in Neural Spike Trains," *Physical Review Letters*, vol. 80, no. 1, pp. 197–200, 1998, ISSN: 0031-9007. DOI: 10.1103/PhysRevLett.80.197. [Online]. Available: https://journals.aps.org/prl/pdf/10.1103/PhysRevLett.80.197%20https://link.aps.org/doi/10.1103/PhysRevE.69.056111%20https://link.aps.org/doi/10.1103/PhysRevLett.80.197.

[146] K. Koch, J. McLean, M. Berry, P. Sterling, V. Balasubramanian, and M. A. Freed, "Efficiency of Information Transmission by Retinal Ganglion Cells," *Current Biology*, vol. 14, no. 17, pp. 1523–1530, 2004, ISSN: 09609822. DOI: 10.1016/j.cub.2004.08.060. [Online]. Available: https://ac.els-cdn.com/S0960982204006566/1-s2.0-S0960982204006566-main.pdf?%7B%5C_%7Dtid=3dfa152e-d580-4890-998f-01b5291f1cd7%7B%5C&%7Dacdnat=1549451672%7B%5C_%7Dda6bc1f8f08d4bee99e6653668b165a0%20https://linkinghub.elsevier.com/retrieve/pii/S0960982204006566.

[147] C. Koch and S. Ullman, "Shifts in selective visual attention: Towards the underlying neural circuitry," in *Matters of intelligence*, Springer, 1987, pp. 115–141.

[148] H. Rebecq, T. Horstschaefer, G. Gallego, and D. Scaramuzza, "EVO: A Geometric Approach to Event-Based 6-DOF Parallel Tracking and Mapping in Real Time," *IEEE Robotics and Automation Letters*, vol. 2, no. 2, pp. 593–600, 2017, ISSN:

2377-3766. DOI: 10.1109/LRA.2016.2645143. [Online]. Available: http://ieeexplore.ieee.org/document/7797445/.

[149] D. H. Hubel and T. N. Wiesel, "Receptive fields, binocular interaction and functional architecture in the cat's visual cortex," *The Journal of physiology*, vol. 160, no. 1, pp. 106–154, 1962.

[150] S. W. Kuffler, "Discharge Patterns And Functional Organization Of Mammalian Retina," *Journal of Neurophysiology*, vol. 16, no. 1, pp. 37–68, 1953. DOI: 10.1152/jn.1953.16.1.37. [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/13035466%20http://www.physiology.org/doi/10.1152/jn.1953.16.1.37.

[151] E. Rubin, *Visuell wahrgenommene figuren [visually perceived patterns]*, 1921.

[152] N. K. Logothetis, D. A. Leopold, and D. L. Sheinberg, "What is rivalling during binocular rivalry?" *Nature*, vol. 380, no. 6575, p. 621, 1996.

[153] E. Craft, H. Schutze, E. Niebur, and R. von der Heydt, "A Neural Model of Figure-Ground Organization," *Journal of Neurophysiology*, vol. 97, no. 6, pp. 4310–4326, 2007. DOI: 10.1152/jn.00203.2007. [Online]. Available: http://jn.physiology.org/cgi/doi/10.1152/jn.00203.2007.

[154] G. Vezzani, U. Pattacini, and L. Natale, "A grasping approach based on superquadric models," in *2017 IEEE International Conference on Robotics and Automation (ICRA)*, IEEE, 2017, pp. 1579–1586.

[155] L. Jamone, A. Bernardino, and J. Santos-Victor, "Benchmarking the grasping capabilities of the icub hand with the ycb object and model set," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 288–294, 2016.

[156] R. Kimchi, Y. Yeshurun, and A. Cohen-Savransky, "Automatic, stimulus-driven attentional capture by objecthood," *Psychonomic Bulletin & Review*, vol. 14, no. 1, pp. 166–172, 2007.

[157] S. Mihalas, Y. Dong, R. Von Der Heydt, and E. Niebur, "Mechanisms of perceptual organization provide auto-zoom and auto-localization for attention to objects," *Proceedings of the National Academy of Sciences*, vol. 108, no. 18, pp. 7583–7588, 2011.

[158] H.-k. Ko, M. Poletti, and M. Rucci, "Microsaccades precisely relocate gaze in a high visual acuity task," *Nature neuroscience*, vol. 13, no. 12, pp. 1549–1553, 2010.

[159] J. Liu, Y. Xiao, Q. Hao, and K. Ghaboosi, "Bio-inspired visual attention in agile sensing for target detection.," *IJSNet*, vol. 5, no. 2, pp. 98–111, 2009.

[160] J. K. Tsotsos and A. Rothenstein, "Computational models of visual attention," *Scholarpedia*, vol. 6, no. 1, p. 6201, 2011.

[161] X. Min, G. Zhai, K. Gu, and X. Yang, "Fixation prediction through multimodal analysis," *ACM Transactions on Multimedia Computing, Communications, and Applications (TOMM)*, vol. 13, no. 1, pp. 1–23, 2016.

[162] C. W. Eriksen and J. D. S. James, "Visual attention within and around the field of focal attention: A zoom lens model," *Perception & psychophysics*, vol. 40, no. 4, pp. 225–240, 1986.

[163] G. R. Mangun, "Neural mechanisms of visual selective attention," *Psychophysiology*, vol. 32, no. 1, pp. 4–18, 1995.

[164] L. Paletta, *Attention in Cognitive Systems. Theories and Systems from an Interdisciplinary Viewpoint: 4th International Workshop on Attention in Cognitive Systems, WAPCV 2007 Hyderabad, India, January 8, 2007 Revised Selected Papers*. Springer Science & Business Media, 2007, vol. 4840.

[165] S. V. Adams, A. D. Rast, C. Patterson, F. Galluppi, K. Brohan, J.-A. Pérez-Carrasco, T. Wennekers, S. Furber, and A. Cangelosi, "Towards real-world neurorobotics: Integrated neuromorphic visual attention," in *International Conference on Neural Information Processing*, Springer, 2014, pp. 563–570.

[166] M. Iacono, G. D'Angelo, A. Glover, V. Tikhanoff, E. Niebur, and C. Bartolozzi, "Proto-object based saliency for event-driven cameras.," in *IROS*, 2019, pp. 805–812.

[167] S. Furber and P. Bogdan, *Spinnaker-a spiking neural network architecture*, 2020.

[168] L. Camunas-Mesa, C. Zamarreño-Ramos, A. Linares-Barranco, A. J. Acosta-Jimenez, T. Serrano-Gotarredona, and B. Linares-Barranco, "An event-driven multi-kernel convolution processor module for event-driven vision sensors," *IEEE Journal of Solid-State Circuits*, vol. 47, no. 2, pp. 504–517, 2011.

[169] C. Lang, T. V. Nguyen, H. Katti, K. Yadati, M. Kankanhalli, and S. Yan, "Depth matters: Influence of depth cues on visual saliency," in *European conference on computer vision*, Springer, 2012, pp. 101–115.

[170] D. A. Burkhardt and P. K. Fahey, "Contrast enhancement and distributed encoding by bipolar cells in the retina," *Journal of Neurophysiology*, vol. 80, no. 3, pp. 1070–1081, 1998.

[171] J. J. Kulikowski, S. Marčelja, and P. O. Bishop, "Theory of spatial position and spatial frequency relations in the receptive fields of simple cells in the visual cortex," *Biological cybernetics*, vol. 43, no. 3, pp. 187–198, 1982.

[172] R. Shapley and V. H. Perry, "Cat and monkey retinal ganglion cells and their visual functional roles," *Trends in Neurosciences*, vol. 9, pp. 229–235, 1986.

[173] T. Sonoda, Y. Okabe, and T. M. Schmidt, "Overlapping morphological and functional properties between m4 and m5 intrinsically photosensitive retinal ganglion cells," *Journal of Comparative Neurology*, vol. 528, no. 6, pp. 1028–1040, 2020.

[174] B. Fischer, "Overlap of receptive field centers and representation of the visual field in the cat's optic tract," *Vision research*, vol. 13, no. 11, pp. 2113–2120, 1973.

[175] M. Chessa, G. Maiello, P. J. Bex, and F. Solari, "A space-variant model for motion interpretation across the visual field," *Journal of vision*, vol. 16, no. 2, pp. 12–12, 2016.

[176] H. Rebecq, D. Gehrig, and D. Scaramuzza, "ESIM: An open event camera simulator," *Conf. on Robotics Learning (CoRL)*, Oct. 2018.

[177] N. Riche, M. Duvinage, M. Mancas, B. Gosselin, and T. Dutoit, "Saliency and human fixations: State-of-the-art and study of comparison metrics," in *Proceedings of the IEEE international conference on computer vision*, 2013, pp. 1153–1160.

[178] C. Willemse and A. Wykowska, "In natural interaction with embodied robots, we prefer it when they follow our gaze: A gaze-contingent mobile eyetracking study," *Philosophical Transactions of the Royal Society B*, vol. 374, no. 1771, p. 20 180 036, 2019.

[179] G. D'Angelo, E. Janotte, T. Schoepe, J. O'Keeffe, M. B. Milde, E. Chicca, and C. Bartolozzi, "Event-based eccentric motion detection exploiting time difference encoding," *Frontiers in neuroscience*, vol. 14, p. 451, 2020.

[180] T. Delbruck, C. Li, R. Graca, and B. Mcreynolds, "Utility and feasibility of a center surround event camera," *arXiv preprint arXiv:2202.13076*, 2022.

[181] B. Hu, R. von der Heydt, and E. Niebur, "Figure-ground organization in natural scenes: Performance of a recurrent neural model compared with neurons of area v2," *Eneuro*, vol. 6, no. 3, 2019.

[182] J. K. Tsotsos, "Analyzing vision at the complexity level," *Behavioral and brain sciences*, vol. 13, no. 3, pp. 423–445, 1990.

[183] A. Yarbus, "Eye movements and vision'plenum press," *New York*, 1967.

[184] D. Walther, L. Itti, M. Riesenhuber, T. Poggio, and C. Koch, "Attentional selection for object recognition—a gentle way," in *International workshop on biologically motivated computer vision*, Springer, 2002, pp. 472–479.

[185] J. M. Wolfe and T. S. Horowitz, "Five factors that guide attention in visual search," *Nature Human Behaviour*, vol. 1, no. 3, pp. 1–8, 2017.

[186] ——, "What attributes guide the deployment of visual attention and how do they do it?" *Nature reviews neuroscience*, vol. 5, no. 6, pp. 495–501, 2004.

[187] D. J. Aks and J. T. Enns, "Visual search for size is influenced by a background texture gradient," *Journal of Experimental Psychology*, vol. 22, no. 6, pp. 1467–1481, 1996.

[188] L. Jansen, S. Onat, and P. König, "Influence of disparity on fixation and saccades in free viewing of natural scenes," *Journal of Vision*, vol. 9, no. 1, pp. 29–29, 2009.

[189] Q. Huynh-Thu and L. Schiatti, "Examination of 3d visual attention in stereoscopic video content," in *Human Vision and Electronic Imaging XVI*, International Society for Optics and Photonics, vol. 7865, 2011, 78650J.

[190] S. May, M. Klodt, E. Rome, and R. Breithaupt, "Gpu-accelerated affordance cueing based on visual attention," in *2007 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2007, pp. 3385–3390.

[191] L. Jamone, E. Ugur, A. Cangelosi, L. Fadiga, A. Bernardino, J. Piater, and J. Santos-Victor, "Affordances in psychology, neuroscience, and robotics: A survey," *IEEE Transactions on Cognitive and Developmental Systems*, vol. 10, no. 1, pp. 4–25, 2016.

[192] K. M. Varadarajan and M. Vincze, "Afrob: The affordance network ontology for robots," in *2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*, IEEE, 2012, pp. 1343–1350.

[193] M. A. Gomez, R. M. Skiba, and J. C. Snow, "Graspable objects grab attention more than images do," *Psychological Science*, vol. 29, no. 2, pp. 206–218, 2018.

[194] A. Pavese and L. J. Buxbaum, "Action matters: The role of action plans and object affordances in selection for action," *Visual cognition*, vol. 9, no. 4-5, pp. 559–590, 2002.

[195] A. Xiong, R. W. Proctor, and H. N. Zelaznik, "Visual salience, not the graspable part of a pictured eating utensil, grabs attention," *Attention, Perception, & Psychophysics*, vol. 81, no. 5, pp. 1454–1463, 2019.

[196] A. Pellicano and F. Binkofski, "The prominent role of perceptual salience in object discrimination: Overt discrimination of graspable side does not activate grasping affordances," *Psychological Research*, vol. 85, no. 3, pp. 1234–1247, 2021.

[197] F. Gabbiani, H. G. Krapp, C. Koch, and G. Laurent, "Multiplicative computation in a visual neuron sensitive to looming," *Nature*, vol. 420, no. 6913, pp. 320–324, 2002, ISSN: 1476-4687. DOI: 10.1038/nature01190. [Online]. Available: https://doi.org/10.1038/nature01190.

[198] S. L. Franconeri and D. J. Simons, "Moving and looming stimuli capture attention," *Perception & Psychophysics*, vol. 65, no. 7, pp. 999–1010, 2003, ISSN: 1532-5962. DOI: 10.3758/BF03194829. [Online]. Available: https://doi.org/10.3758/BF03194829.

[199] M. Yilmaz and M. Meister, "Rapid innate defensive responses of mice to looming visual stimuli," en, *Curr Biol*, vol. 23, no. 20, pp. 2011–2015, Oct. 2013.

[200] Y. Yu, G. K. Mann, and R. G. Gosine, "An object-based visual attention model for robotic applications," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 40, no. 5, pp. 1398–1412, 2010.

[201] H.-k. Ko, D. M. Snodderly, and M. Poletti, "Eye movements between saccades: Measuring ocular drift and tremor," *Vision research*, vol. 122, pp. 93–104, 2016.

[202] A. Glover, V. Vasco, M. Iacono, and C. Bartolozzi, "The event-driven Software Library for YARP — With Algorithms and iCub Applications," *Frontiers in Robotics and AI*, vol. 4, p. 73, 2018. DOI: 10.3389/frobt.2017.00073.

[203] Y.-D. Zhu and N. Qian, "Binocular receptive field models, disparity tuning, and characteristic disparity," *Neural computation*, vol. 8, no. 8, pp. 1611–1641, 1996.

[204] D. Marr and T. Poggio, "Cooperative computation of stereo disparity," *Science*, vol. 194, no. 4262, pp. 283–287, 1976.

[205] G. Dikov, M. Firouzi, F. Röhrbein, J. Conradt, and C. Richter, "Spiking cooperative stereo-matching at 2 ms latency with neuromorphic hardware," in *Conference on Biomimetic and Biohybrid Systems*, Springer, 2017, pp. 119–137.

[206] E. Piatkowska, A. Belbachir, and M. Gelautz, "Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach," in *Proceedings of the IEEE International Conference on Computer Vision Workshops*, 2013, pp. 45–50.

[207] B. W. Knight, "Dynamics of encoding in a population of neurons," *The Journal of general physiology*, vol. 59, no. 6, pp. 734–766, 1972.

[208] J. Harel, C. Koch, and P. Perona, "Graph-based visual saliency," *Advances in neural information processing systems*, vol. 19, 2006.

[209] M. E. Nelson and M. A. MacIver, "Sensory acquisition in active sensing systems," *Journal of Comparative Physiology A*, vol. 192, no. 6, pp. 573–586, 2006.

[210] J. H. Maunsell and E. P. Cook, "The role of attention in visual processing," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 357, no. 1424, pp. 1063–1072, 2002.

[211] B. D. Lucas, T. Kanade, *et al.*, "An iterative image registration technique with an application to stereo vision," 1981.

[212] R. Benosman, S.-H. Ieng, C. Clercq, C. Bartolozzi, and M. Srinivasan, "Asynchronous frameless event-based optical flow," *Neural Networks*, vol. 27, pp. 32–37, 2012.

[213] T. Brosch, S. Tschechne, and H. Neumann, "On event-based optical flow detection," *Frontiers in neuroscience*, vol. 9, p. 137, 2015.

[214] M. B. Milde, O. J. Bertrand, R. Benosmanz, M. Egelhaaf, and E. Chicca, "Bioinspired event-driven collision avoidance algorithm based on optic flow," in *2015 International Conference on Event-based Control, Communication, and Signal Processing (EBCCSP)*, IEEE, 2015, pp. 1–7.

[215] T. Horiuchi, J. Lazzaro, A. Moore, and C. Koch, "A delay-line based motion detection chip," in *Advances in neural information processing systems*, 1991, pp. 406–412.

[216] A. Borst, J. Haag, and D. F. Reiff, "Fly motion vision," *Annual review of neuroscience*, vol. 33, pp. 49–70, 2010.

[217] M. S. Maisak, J. Haag, G. Ammer, E. Serbe, M. Meier, A. Leonhardt, T. Schilling, A. Bahl, G. M. Rubin, A. Nern, *et al.*, "A directional tuning map of drosophila elementary motion detectors," *Nature*, vol. 500, no. 7461, pp. 212–216, 2013.

[218] A. S. Mauss, M. Meier, E. Serbe, and A. Borst, "Optogenetic and pharmacologic dissection of feedforward inhibition in drosophila motion vision," *Journal of Neuroscience*, vol. 34, no. 6, pp. 2254–2263, 2014.

[219] A. Borst and M. Helmstaedter, "Common circuit design in fly and mammalian motion vision," *Nature Neuroscience*, vol. 18, no. 8, pp. 1067–1076, 2015.

[220] J. A. Strother, S.-T. Wu, A. M. Wong, A. Nern, E. M. Rogers, J. Q. Le, G. M. Rubin, and M. B. Reiser, "The emergence of directional selectivity in the visual motion pathway of drosophila," *Neuron*, vol. 94, no. 1, pp. 168–182, 2017.

[221] R. C. Nelson and J. Aloimonos, "Obstacle avoidance using flow field divergence," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 11, no. 10, pp. 1102–1106, 1989.

[222] A. Gelbukh, F. C. Espinoza, and S. N. Galicia-Haro, *Human-Inspired Computing and its Applications: 13th Mexican International Conference on Artificial Intelligence, MICAI2014, Tuxtla Gutiérrez, Mexico, November 16-22, 2014. Proceedings*. Springer, 2014, vol. 8856.

[223] R. Benosman, C. Clercq, X. Lagorce, S.-H. Ieng, and C. Bartolozzi, "Event-based visual flow," *Neural Networks and Learning Systems, IEEE Transactions on*, vol. 25, no. 2, pp. 407–417, 2014.

[224] G. Gallego, M. Gehrig, and D. Scaramuzza, "Focus is all you need: Loss functions for event-based vision," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 12 280–12 289.

[225] A. Mitrokhin, C. Fermüller, C. Parameshwara, and Y. Aloimonos, "Event-based moving object detection and tracking," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2018, pp. 1–9.

[226] M. Giulioni, X. Lagorce, F. Galluppi, and R. B. Benosman, "Event-based computation of motion flow on a neuromorphic analog neural platform," *Frontiers in neuroscience*, vol. 10, p. 35, 2016.

[227] G. Haessig, A. Cassidy, R. Alvarez, R. Benosman, and G. Orchard, "Spiking optical flow for event-based sensors using ibm's truenorth neurosynaptic system," *IEEE transactions on biomedical circuits and systems*, vol. 12, no. 4, pp. 860–870, 2018.

[228] T. Schoepe, D. Gutierrez-Galan, J. Dominguez-Morales, A. Jimenez-Fernandez, A. Linares-Barranco, and E. Chicca, "Neuromorphic sensory integration for combining sound source localization and collision avoidance," *2019 IEEE Biomedical Circuits and Systems Conference (BioCAS)*, pp. 1–4, 2019, ISSN: 2163-4025. DOI: 10.1109/BIOCAS.2019.8919202.

[229] A. Traschütz, W. Zinke, and D. Wegener, "Speed change detection in foveal and peripheral vision," *Vision Research*, vol. 72, pp. 1–13, 2012.

[230] J. Freeman and E. P. Simoncelli, "Metamers of the ventral stream," *Nature neuroscience*, vol. 14, no. 9, p. 1195, 2011.

[231] J. Wurbs, E. Mingolla, and A. Yazdanbakhsh, "Modeling a space-variant cortical representation for apparent motion," *Journal of vision*, vol. 13, no. 10, pp. 2–2, 2013.

[232] B. M. Harvey and S. O. Dumoulin, "The relationship between cortical magnification factor and population receptive field size in human visual cortex: Constancies in cortical architecture," *Journal of Neuroscience*, vol. 31, no. 38, pp. 13 604–13 612, 2011.

[233] F. M. Panerai, C. Capurro, and G. Sandini, "Space-variant vision for an active camera mount," in *Visual Information Processing IV*, International Society for Optics and Photonics, vol. 2488, 1995, pp. 284–296.

[234] A. Bernardino and J. Santos-Victor, "Binocular tracking: Integrating perception and control," *IEEE Transactions on Robotics and Automation*, vol. 15, no. 6, pp. 1080–1094, 1999.

[235] B. Ramesh, H. Yang, G. M. Orchard, N. A. Le Thi, S. Zhang, and C. Xiang, "Dart: Distribution aware retinal transform for event-based cameras," *IEEE transactions on pattern analysis and machine intelligence*, 2019.

[236] H. Wässle and H. Riemann, "The mosaic of nerve cells in the mammalian retina," *Proceedings of the Royal Society of London. Series B. Biological Sciences*, vol. 200, no. 1141, pp. 441–461, 1978.

[237] S. H. Devries and D. A. Baylor, "Mosaic arrangement of ganglion cell receptive fields in rabbit retina," *Journal of neurophysiology*, vol. 78, no. 4, pp. 2048–2060, 1997.

[238] I. Murray, F. MacCana, and J. Kulikowski, "Contribution of two movement detecting mechanisms to central and peripheral vision," *Vision Research*, vol. 23, no. 2, pp. 151–159, 1983.

[239] R. A. Abrams and S. E. Christ, "The onset of receding motion captures attention: Comment on franconeri and simons (2003)," *Perception & Psychophysics*, vol. 67, no. 2, pp. 219–223, 2005.

[240] C. J. Howard and A. O. Holcombe, "Unexpected changes in direction of motion attract attention," *Attention, Perception, & Psychophysics*, vol. 72, no. 8, pp. 2087–2095, 2010.

[241] A. P. Hillstrom and S. Yantis, "Visual motion and attentional capture," *Perception & psychophysics*, vol. 55, no. 4, pp. 399–411, 1994.

[242] S. J. Van Albada, A. G. Rowley, J. Senk, M. Hopkins, M. Schmidt, A. B. Stokes, D. R. Lester, M. Diesmann, and S. B. Furber, "Performance comparison of the digital neuromorphic hardware spinnaker and the neural network simulation software nest for a full-scale cortical microcircuit model," *Frontiers in neuroscience*, vol. 12, p. 291, 2018.

[243] C. Pehle, S. Billaudelle, B. Cramer, J. Kaiser, K. Schreiber, Y. Stradmann, J. Weis, A. Leibfried, E. Müller, and J. Schemmel, "The brainscales-2 accelerated neuromorphic system with hybrid plasticity," *Frontiers in Neuroscience*, vol. 16, 2022.

[244] R. Jerath, S. M. Cearley, V. A. Barnes, and E. Nixon-Shapiro, "How lateral inhibition and fast retinogeniculo-cortical oscillations create vision: A new hypothesis," *Medical Hypotheses*, vol. 96, pp. 20–29, 2016, ISSN: 0306-9877. DOI: https://doi.org/10.1016/j.mehy.2016.09.015. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0306987716306168.

[245] W. Gerstner, W. M. Kistler, R. Naud, and L. Paninski, *Neuronal Dynamics: From Single Neurons to Networks and Models of Cognition*. Cambridge University Press, 2014. DOI: 10.1017/CBO9781107447615.

Appendices

Appendix A

Event-driven Proto-object based saliency in 3D space to attract a robot's attention –Supplementary Material–

Multimedia Material

A supplementary video summarizing proposed methodology and results can be found at `https://zenodo.org/record/5091539/files/evProtoDepth.mp4?download=1`

Event-based proto-object model structure and parameters

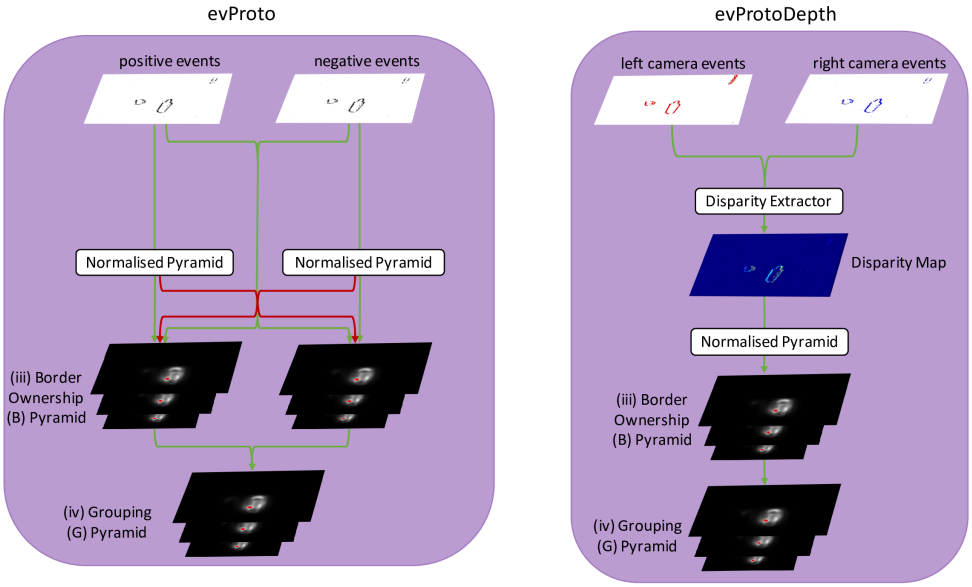


Figure A.1. Schematic representation of the evProto (left) and evProtoDepth (right) models. Both models show size invariance due to the 5 levels of the Normalised Pyramid. In the evProto model, positive and negative events go into the two core stages of the model: Border Ownership and Grouping. In the case of the evProtoDepth, the binocular events are fed into the disparity Extractor producing a Disparity Map. The Border Ownership takes in input the normalised outcome from the Pyramid fed with the disparity events.

Parameter	Value
R_0	10
ρ	0.2
$Pyramid\ levels$	5
$Orientations$	0°, 45°, 90°, 135°

Table A.1. Parameters used in the 2D proto-object model [166] and the proposed evProtoDepth model. R_0 is the radius of the filter, ρ determines the arc length of active pixels in the kernel allowing to change the convexity of the kernel, $Pyramid\ levels$ the number of scales and the $Orientations$ of the Von Mises filters.

A schematic diagram of the event-driven proto-object models in both 2D (evProto [166]) and 3D (our proposed evProtoDepth) is depicted in Fig A.1. The parameters used in the proto-object model are listed in Table A.1.

Disparity computation model

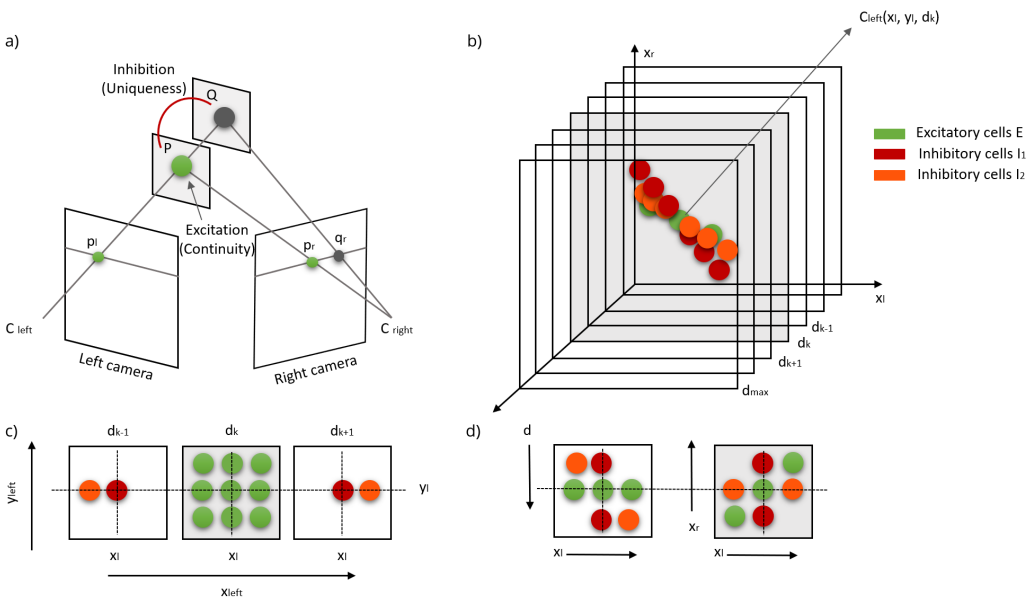


Figure A.2. Asynchronous Cooperative Stereo Network for event-based disparity estimation[11]. a) Representation of the stereopsis problem describing continuity and uniqueness constraints b) three-dimensional view of the Excitation and Inhibition Network. The axes refer to the horizontal coordinates of the two cameras and the disparity c), d) Excitation and Inhibition Network on a two-dimensional plane: c)x_l, y_l axes and d) $disparity$, x_l & x_l, x_r axes.

We computed disparity of the scene on a per-event basis using a cooperative network that employs time correspondence between a stereo event-pair and imposes the disparity uniqueness and continuity conditions proposed by Marr and Poggio[204] to model a correspondence belief map. Inspired from Firouzi et al.[11], we used a 3D array-based representation of the network, called an activity map C, which gets updated with each

incoming input event. Each element (cell) of this array abstracts a computational neuron in the Spiking Neural Network, where each correspondence neuron spikes during simultaneous triggering of events in its associated left and right pixels. At any time instant, the current state of the activity map represents the disparity of the present input scene. Each incoming event from the left or right camera gets processed asynchronously in the network without any explicit synchronization between them. A schematic diagram of the cooperative network used is shown in Fig A.2.

Each incoming event was remapped to its pixels coordinate using pre-calibrated stereo camera parameters to ensure that the corresponding left and right events have the same y coordinates. The stereo correspondence search thus gets simplified to a single row scan due to the inherent epipolar constraints. We define disparity as $d = x_l - x_r$, where x_l and x_r are rectified events from the left and right cameras respectively. Since the rectified image planes are parallel, we only have positive disparities. Due to the asymmetric nature of the disparity computation, the left and right events are processed non-identically. We present here the relevant computations performed on an incoming left event. The corresponding operations to be performed on a right event then follows by changing the signs for disparity equations as the direction of matching switches. Furthermore, the activity map and hence the disparity map, is represented with the reference of the left camera frame. It is a matter of arbitrary choice, and each pixel coordinate only needs to be horizontally shifted by its computed disparity when a right reference frame is used.

We consider an input rectified event $E_l = (P_l, t_l)$ that is generated from the left sensor at time t_l and pixel location $P_l = (x_l, y_l)$. The cooperative network computes the best disparity value for this event. Each activity cell C_{x_l,y_l,d_k} encodes the belief of the system about whether x_l from the left event and $x_r = x_l - d_k$ from the right event are true stereo correspondences, in the form of activity. Consequently, each cell encodes the validity of the disparity value d_k for event E_l. The size of the matrix is thus $M \times N \times d_{max}$, where $M \times N$ is the sensor dimension.

Due to epipolar constraints, the set of possible corresponding pixels in the right image are:

$$S_l = \{(x_r, y_r) \mid x_l - d_{max} \leq x_r \leq x_l,\ y_r = y_l\} \tag{A.1}$$

where d_{max} is an algorithmic parameter that determines the maximum detectable disparity. Each element of this correspondence set represents a layer d_k in the activity map for P_l. Therefore, a cell C_{x_l,y_l,d_k} represents a candidate correspondence in S_l. For a single incoming event, we thus compute the activity related to each candidate correspondence in

S_l. The cooperative network evaluates the legitimacy of each candidate in S_l for its true correspondence to P_l, using a Winner-Takes-All mechanism. The winner disparity value d_{WTA}, representing the d_k layer with the maximum activity value (above a predefined threshold θ), is the final computed disparity for the input event E_l.

$$d_{WTA}(E_l) = \arg\max_{d_k} \left\{ C_{x_l,y_l,d_k} \mid C_{x_l,y_l,d_k} \geq \theta \right\} \tag{A.2}$$

The activity for each cell is computed using time-weighted excitatory and inhibitory connections from previously activated cells in the network. These connections are determined according to the continuity and uniqueness disparity constraints respectively.

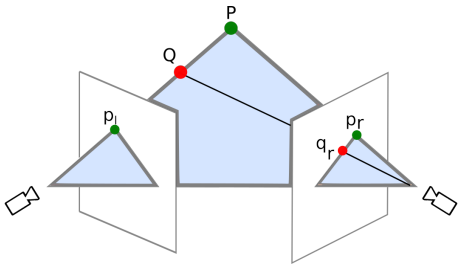


Figure A.3. Event correspondence

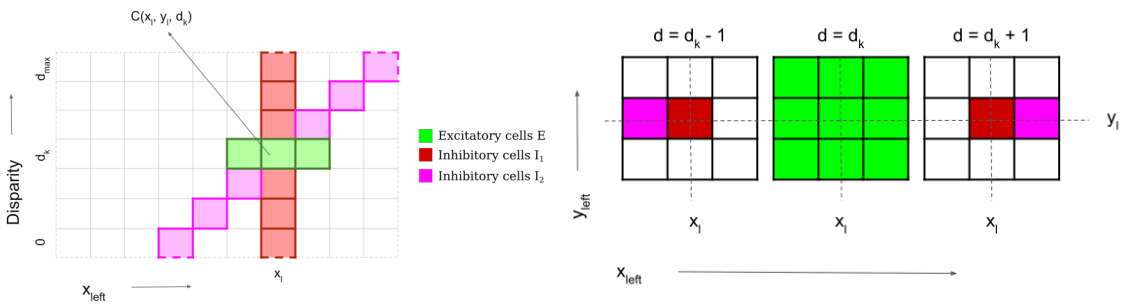


Figure A.4. C_{x_l,y_l,d_k} depicted in cross-sectional views of the activity map, along with its excitatory and inhibitory sets, given by equations A.3, A.4 and A.5, with parameters $d_{max} = 6$ and $r = 1$. Top: cross-sectional view at vertical layer y_l; Bottom: cross-sectional views at disparity layers $d_k - 1$, d_k and $d_k + 1$.

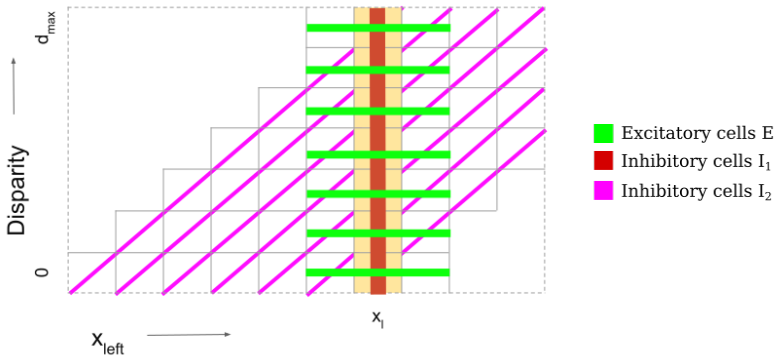


Figure A.5. Cross-section of activity map C at layer y_l, showing all excitatory and inhibitory sets for an input left event e_l at pixel $P_l = (x_l, y_l)$, with parameters $d_{max} = 6$ and $r = 1$. Each yellow element in the column x_l represents an activity cell $C_{x_l,y_l,d}$. The lines of the excitatory and inhibitory sets for each cell intersect at its center. Inhibition set I_1 is nearly same (all cells along the x_l column except the cell in focus) for all candidate correspondences.

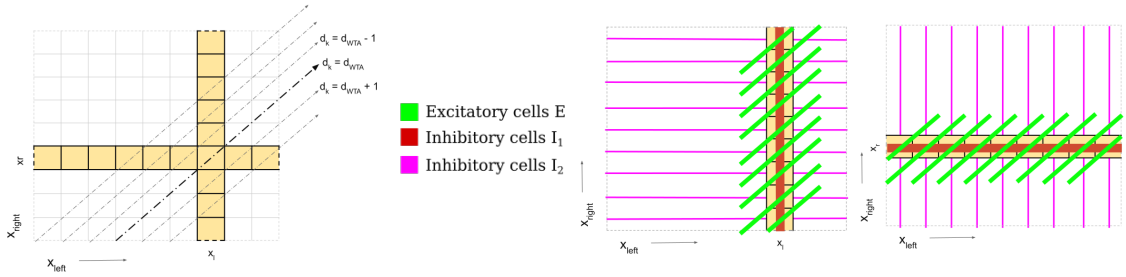


Figure A.6. Network computations for matching left and right event pair with pixel coordinates (x_l, y) and (x_r, y) respectively, with parameters $d_{max} = 6$ and $r = 1$. Top-left: Representation of estimated true disparity d_{WTA} in left-right x-coordinate correspondence map; Bottom-left: All network excitations and inhibitions for input left event at (x_l, y); Bottom-right: All network excitations and inhibitions for input right event at (x_r, y).

Excitatory connections

To enforce the within-disparity continuity constraint, cells in the same disparity layer surrounding the firing pixel should potentiate each other. This leads to more contiguous regions in the disparity map based on the reasoning that pixels in the close neighbourhood should possess similar disparity values. Events generated by extended objects in the spatial and temporal vicinity thus lead to more accurate disparity values. Therefore, for each element in S_l, the set of excitatory connections are defined as:

$$E(C_{x_l,y_l,d_k}) = \{C_{x',y',d_k} \mid |x' - x_l| \leq r,\, |y' - y_l| \leq r\} \tag{A.3}$$

where r is the size of the neighborhood which we consider for excitation.

Inhibitory connections

To enforce the cross-disparity uniqueness conditions, cells in the same epipolar line contributing to other disparities should inhibit the current disparity belief evaluation. For each candidate correspondence in S_l, we use two kinds of inhibitory connections in the network.

The first set of inhibitory connections are defined as

$$I_1(C_{x_l,y_l,d_k}) = \{C_{x',y',d} \mid 0 \leq d \leq d_{max},\, d \neq d_k,\, x' = x_l,\, y' = y_l\} \tag{A.4}$$

This set of inhibitory cells topologically represent correspondence between the current cell C_{x_l,y_l,d_k} and all other pixels lying on the conjugate epipolar line of P_l in the right image. This means that for a left event, a candidate disparity that has already been assigned a high belief by the network, inhibits all other possible disparities for that event. Lateral inhibition like this is present throughout the human vision system [244]. It helps to reduce false positive matches in noisy environments. A false positive matching scenario is depicted in figure A.3 – if p_l on the left image is a retinal projection of the point object P, there are two possible matches in the right retina, p_r or q_r. However, since just one of the candidates can be chosen, the correspondence p_l - p_r should inhibit p_l - q_r so that it does not lead to the false positive 3D point Q.

The second set of inhibitory connections is defined as:

$$I_2(C_{x_l,y_l,d_k}) = \{C_{x',y',d} \mid 0 \leq d \leq d_{max},\, d \neq d_k,\, x' = x_l - d_k + d,\, y' = y_l\} \quad \text{(A.5)}$$

This layer of inhibition, as proposed in [11], is used to further reduce false matches by enforcing that a candidate right pixel should contribute to only one stereo correspondence. For each disparity level d_k, it may happen that the corresponding right pixel $(x_r = x_l - d_k,\, y_r = y_l)$ in S_l has already contributed to a stereo match in an earlier iteration of the algorithm. We thus inhibit the current correspondence belief C_{x_l,y_l,d_k}, using the summation of cell activities (x_r, y_r) might have contributed to. Since the C is formulated with reference to the left image frame, x_r is actually represented as $x' = x_r + d$ in the activity map. This extra layer of inhibition resolves disparity ambiguity in the scene caused by multiple bodies, therefore producing more precise disparity maps.

Figure A.4 illustrates the excitatory set E (in green), as well as inhibitory sets I_1 (in red) and I_2 (in magenta) for a single activity cell C_{x_l,y_l,d_k}. Both (x_{left}-disparity) and (x_{left}-y_{left}) cross-sections of a section of the 3D activity map C are shown. These sets are computed for each candidate correspondence element in S_l. Thus, computations for an input left event e_l with pixel coordinates $P_l = (x_l, y_l)$ are affected by multiple excitatory and inhibitory sets as depicted in figure A.5.

Temporal Correspondence for Activity computation

Events are triggered by the ATIS cameras whenever there is change is illumination in the input scene. Thus, events originating from the left and right cameras due to the same source are generated around the same time. The temporal proximity of corresponding events thus helps in better stereo matching. Ideally, temporal coincidence should represent pixel correspondence on an event level. However, input stimuli with noise and multiple extended objects generate a lot of events around the same time, thus the timing information encoded in the events are not perfect due to jitter in latency of the acquisition system from the left and right cameras. Furthermore, the order of generated events from each camera may also be incoherent. The cooperative network we use ensures that the disparity estimation works even when timing information is not precise.

To ensure that temporally close events have higher probability to correspond to each other, a simplified abstraction of Leaky Integrate and Fire (LIF) model [245] is used to

model the internal dynamics of each activity cell. An activation time is maintained for each cell in the activity map. Every time a cell in the network $C_{x,y,d}$ gets activated due to an incoming input event, we update its activation time $t_{x,y,d}$. We use a temporal kernel W that weights the contribution of each interacting cell (excitatory and inhibitory) based on how far ago in time they were activated, with respect to the current event time t_l. It is defined as follows:

$$W^{t_l}_{x,y,d} = \frac{1}{1 + \beta(t_l - t_{x,y,d})} \tag{A.6}$$

where t_l is the current activation time, and $t_{x,y,d}$ is the time when the activity of cell $C_{x,y,d}$ was last activated.

Therefore, for an incoming left event E_l, activities of all network cells, each corresponding to a disparity layer, are computed in a single pass and stored in temporary one-dimensional array. Using the temporal kernel W, excitation set E, and inhibition sets I_1 and I_2, the activity of each cell C_{x_l,y_l,d_k} is computed as:

$$C_{x_l,y_l,d_k} = \sigma\left(\sum_{x',y',d' \epsilon\, E} W^{t_l}_{x',y',d'} C_{x',y',d'} \; - \; \alpha \sum_{x',y',d' \epsilon\, I_1 \cup I_2} W^{t_l}_{x',y',d'}\, C_{x',y',d'} \right) \tag{A.7}$$

where the sigmoid function $\sigma(k) = \frac{1}{1+e^{-k}}$ is used to normalize the activity to values between 0 and 1. Using equation A.2, we estimate the disparity value for E_l. The activity values and their respective trigger times are finally updated inside the network in a single pass.

When two events e_l and e_r from the left and right cameras are generated close in time, the network computes their disparity as $d_{WTA} = x_l - x_r$, where x_l and x_r are respective horizontal pixel coordinates of the events. This is ensured by appropriately weighting the inhibitory and excitatory connections in the network, using the parameter α. Figure A.6 illustrates the computations performed by the network for these two corresponding events. In this figure, all connections are plotted in the domain of x_{left} - x_{right} correspondence maps for symmetrical representation. The candidate matching cells for both events (shown in yellow) lie along rows and columns of the correspondence map. Layers of constant disparity are shown with dashed arrows. The figure explains how the cooperative network excites disparity in neighbouring regions, and inhibits disparities along epipolar lines. However, for ease of implementation, we map all computations in the x -

disparity domain, like in figure A.5. This leads to efficient traversal across disparity layers represented by straight lines along contiguous 1-D array elements.

Parameter	Value
r	3
θ	0.4
α	0.3
β	0.0001
d_{max}	45

Table A.2. Parameters used for disparity computation using the asynchronous cooperative network. Excitatory neighborhood r tunes the smoothness of the disparity maps. Activation function threshold θ can be used to adjust the desired trade-off between noise and sparsity of output – high values will filter out noisy predictions with low activity but may also remove valid estimates. Inhibitory factor α tunes the strength of inhibition during cooperation. Slope of temporal correlation kernel β can adjust the temporal sensitivity of the cells to input events – higher value means faster dynamics and sharper temporal sensitivity to the upcoming events. The number of disparity levels d_{max} is used to modulate the precision of disparity estimation at the cost of increased computational overheads.

The empirically chosen parameters of disparity extractor are listed in Table A.2.

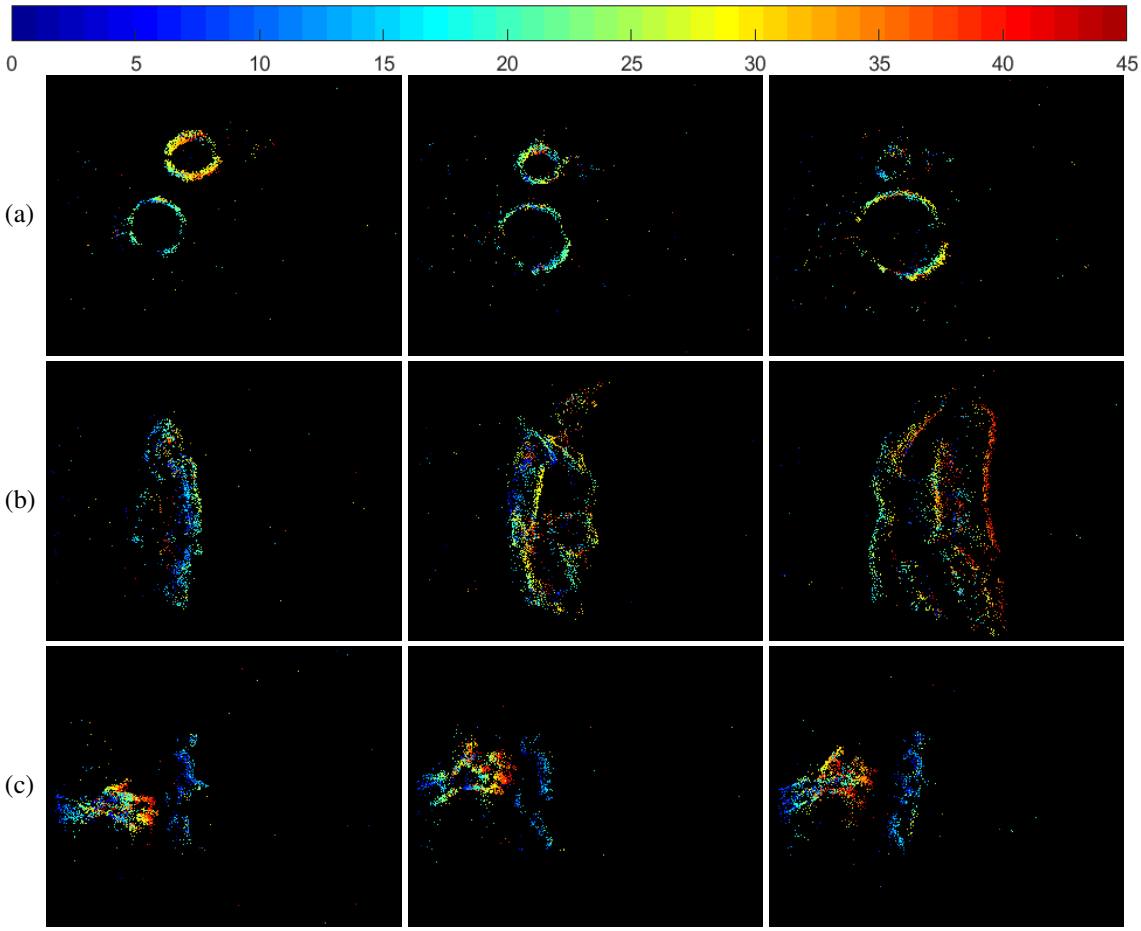


Figure A.7. Estimated disparity for various instances of three sequences: a Two paddles of different sizes moving towards and away from the cameras, along the depth axis of the robot's cameras. Pixels where events occur during the 100ms time-window are colour-coded as per the computed disparity (depth). The object stimuli switch colours as their relative depth changes. b Same as (a) but for a person with at varying distances between 30 cm and 210 cm from the robot. c Scene showing two persons at different depths from the robot. Person 1 (mainly yellow and red pixes) is waving a hand at 30 cm depth. Person 2 (mainly blue) is walking horizontally across the scene at ≈210 cm depth. Pixel colours represent depth, as in (a) and (b), and remain constant for motion in the same depth plane.

Saliency Benchmarking with NUS-3D dataset

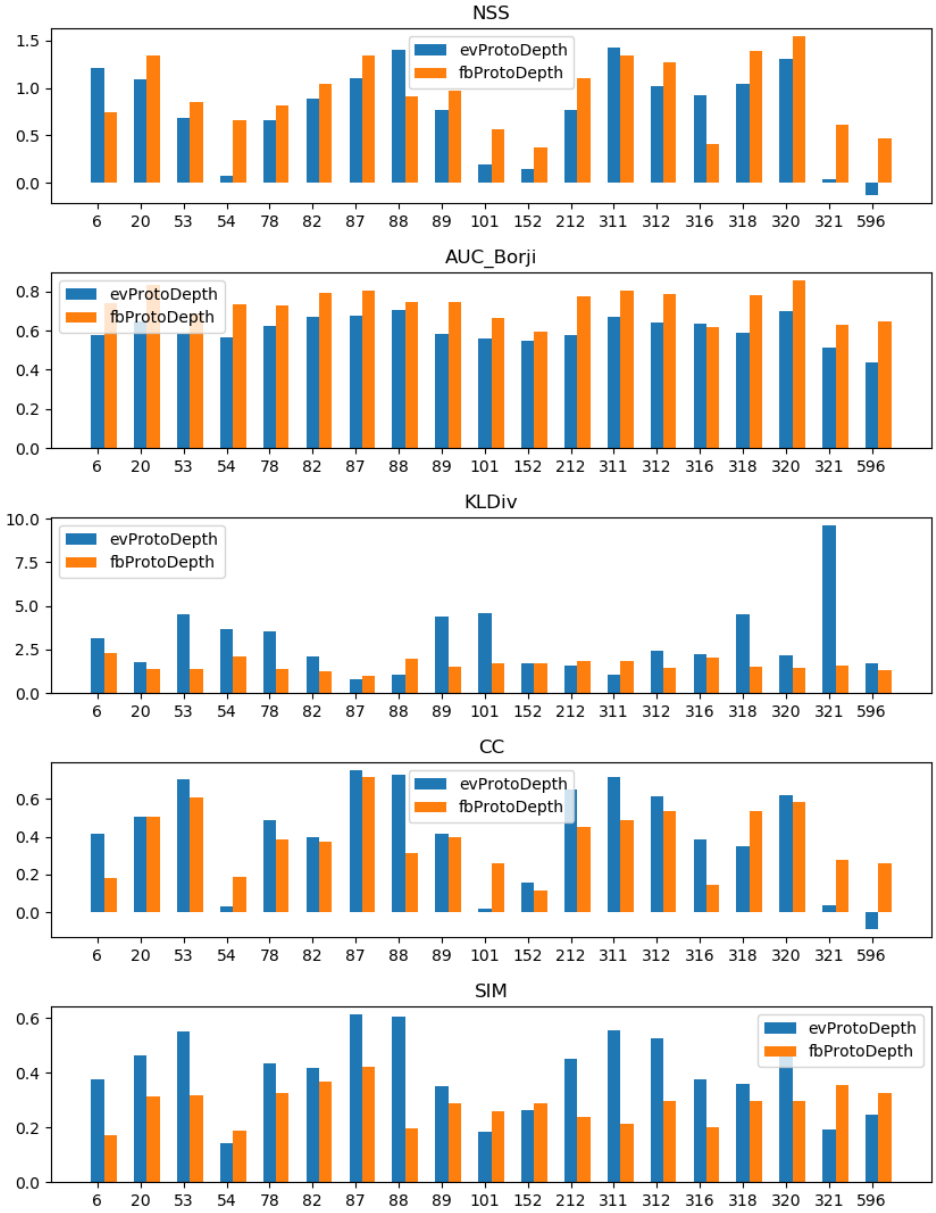


Figure A.8. Quantitative evaluation of saliency maps generated by fbProtoDepth [88] and evProtoDepth using the MIT saliency metrics Normalized Scanpath Saliency (NSS), Area under the ROC Curve (AUC-Borji), Kullback-Leibler Divergence (KLDiv), Pearson's Correlation Coefficent (CC) and Similarity (SIM) [35], [137]–[139] on a subset of the NUS3D dataset where human eyes were fixated is mostly on the nearest object of the scene. The subset comprises all cases among the NUS-3D dataset where the cross-correlation between the ground truth 3D fixation and inverse of ground truth depth ≥ 0.5. The x-axis depicts the image number as present in the NUS-3D dataset. For all metrics except KLDiv, larger value is better.

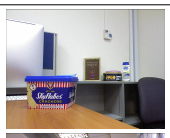
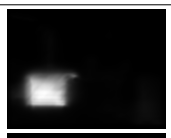
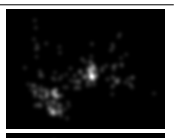
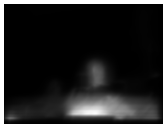
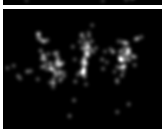
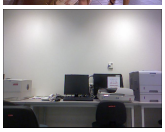
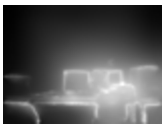
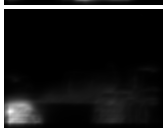
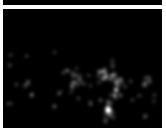
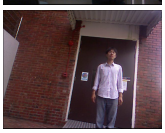
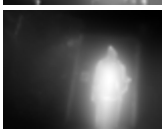
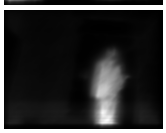
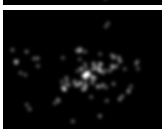
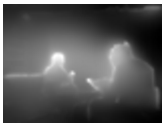
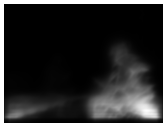
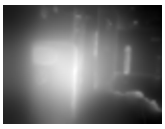
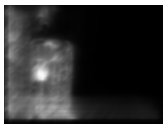
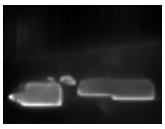
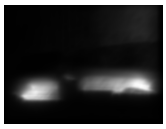
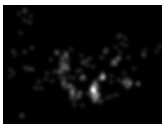
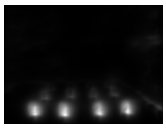
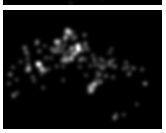
Image #	RGB image	Saliency Map: fbProtoDepth [88]	Saliency Metrics: fbProtoDepth [88]	Saliency Map: evProtoDepth	Saliency Metrics: evProtoDepth	Ground truth 3D fixations
156			**NSS = 0.808** **AUC-Borji = 0.708** **KLDiv = 1.343** **CC = 0.387** **Sim = 0.362**		NSS = 0.655 AUC-Borji = 0.6 KLDiv = 3.215 CC = 0.364 Sim = 0.361	
188			**NSS = 0.856** **AUC-Borji = 0.74** **KLDiv = 1.176** **CC = 0.454** **Sim = 0.384**		NSS = 0.357 AUC-Borji = 0.602 KLDiv = 5.086 CC = 0.116 Sim = 0.303	
208			**NSS = 0.693** **AUC-Borji = 0.704** **KLDiv = 1.516** **CC = 0.283** **Sim = 0.285**		NSS = 0.339 AUC-Borji = 0.606 KLDiv = 3.622 CC = 0.171 Sim = 0.258	
250			**NSS = 0.935** **AUC-Borji = 0.74** **KLDiv = 1.563** **CC = 0.372** **Sim = 0.285**		NSS = 0.15 AUC-Borji = 0.531 KLDiv = 3.3 CC = 0.054 Sim = 0.198	
300			**NSS = 0.79** **AUC-Borji = 0.705** **KLDiv = 1.556** **CC = 0.331** **Sim = 0.292**		NSS = 0.482 AUC-Borji = 0.588 KLDiv = 2.319 CC = 0.172 Sim = 0.252	
350			**NSS = 0.808** **AUC-Borji = 0.713** **KLDiv = 1.388** **CC = 0.467** **Sim = 0.318**		NSS = -0.031 AUC-Borji = 0.503 KLDiv = 6.676 CC = -0.051 Sim = 0.139	
410			**NSS = 0.557** **AUC-Borji = 0.657** **KLDiv = 1.379** **CC = 0.271** **Sim = 0.325**		NSS = -0.057 AUC-Borji = 0.479 KLDiv = 12.357 CC = -0.037 Sim = 0.192	
500			**NSS = 0.282** **AUC-Borji = 0.636** **KLDiv = 1.466** **CC = 0.203** **Sim = 0.312**		NSS = 0.236 AUC-Borji = 0.534 KLDiv = 2.157 CC = 0.15 Sim = 0.251	
536			**NSS = 1.099** **AUC-Borji = 0.81** **KLDiv = 1.809** **CC = 0.372** **Sim = 0.248**		NSS = 0.602 AUC-Borji = 0.607 KLDiv = 2.335 CC = 0.13 Sim = 0.236	
569			**NSS = 0.252** **AUC-Borji = 0.565** **KLDiv = 1.846** **CC = 0.337** **Sim = 0.251**		NSS = -0.218 AUC-Borji = 0.431 KLDiv = 6.576 CC = -0.138 Sim = 0.096	

Figure A.9. Example scenes from the publicly available NUS-3D dataset [169] where the frame-based fbProtoDepth model [88], which includes colour opponency as well as orientation information channels, quantitatively outperforms the event-based evProtoDepth model in all the MIT saliency benchmark metrics. Here, the ground truth fixations are not confined to the nearest "object" in the scene. Better values are bold-faced. The event-based model generates more localised saliency maps that are better at precisely selecting the nearest "object" since it mainly relies on depth and Gestalt cues from high contrast edges.

Robot experiments

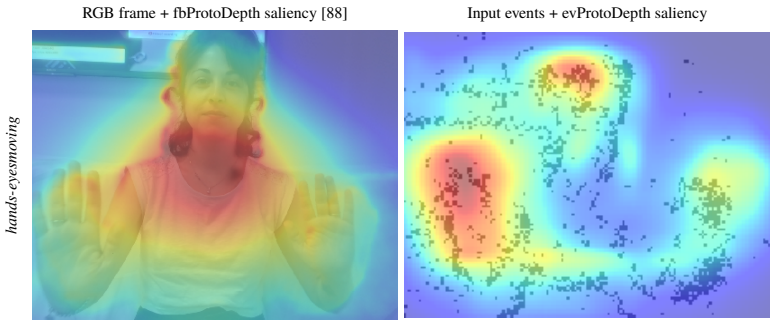


Figure A.10. Comparison of saliency map selectivity between fbProtoDepth and evProtoDepth models running on the robot. Each plot shows the 2D histogram of accumulated saliency maps over all the frames of the *clutter* and *hands-eyesmoving* datasets. The frame-based model receives RGB and depth input from the RealSense camera, and the event-driven models receive input from the stereo event cameras. The two cameras have different field of views, and hence are not spatially aligned.

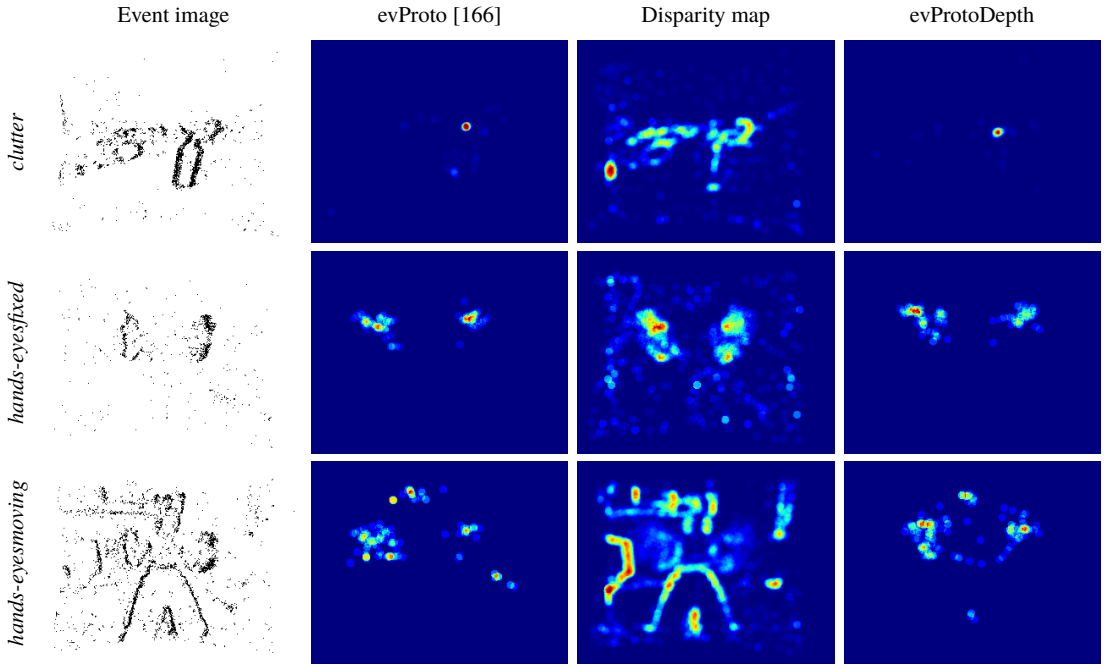


Figure A.11. Contribution of both disparity and proto-object modelling in selecting the nearest (most salient) "object" on the *clutter* and *hands* datasets. Each plot depicts the 2D histogram of peak saliency pixel over all the frames of a dataset. For the disparity-only setup in Column 3, the peaks of the raw disparity map also includes non-salient regions of the scene, hence using disparity alone is not suitable for stable object selection. For the proto-object models in Column 2 and 4, the saliency peaks are more precisely concentrated at suitable proto-objects. The combination of disparity and proto-object modelling in evProtoDepth thus generates more robust and precise peak saliency locations than the individual 2D evProto model or the disparity map. While the disparity information improves selectivity of the saliency model, the proto-object model acts as filter to isolate high-level "proto-objects" from noisy depth maps.

Appendix B

Event-based eccentric motion detection exploiting time difference encoding –Supplementary Material–

B.1 Experiments real-world data

Final experiments were conducted using recorded data as input to understand the behaviour of the model to a real scenario. We analysed various recordings showing a black bar moving from left to right on a white canvas. The datasets differ in terms of stimulus speed. Given that we want to simulate a real robotic scenario we manually moved the bar in front of the camera. However, as the bar was moved manually, a constant velocity could not be guaranteed. Hence, this is not a comparison with the simulated input experiments because is out of the scope of this analysis. The aim of these recordings was to show the real-world response of the model and its robustness in detecting the correct speed regardless to the noise while decreasing the incoming events from the cameras thanks to the eccentric down-sampling.

B.2 Experiments real-world data results

Figure B.1 shows the response of the sEMD with eccentric down-sampling for real-world input data from the ATIS camera. Three different stimulus speeds are visualised, ranging from slow B.1 a), to medium B.1 b) and fast B.1 c). As the stimuli were moved manually, no more reliable assertion of the velocities can be made. The observed responses showed the same trend observed for the simulated data. An increase of stimulus speed causes a shift of the area of highest response from the fovea to the periphery. However, unlike observed with the simulated data, the mean firing rate seems to decrease instead

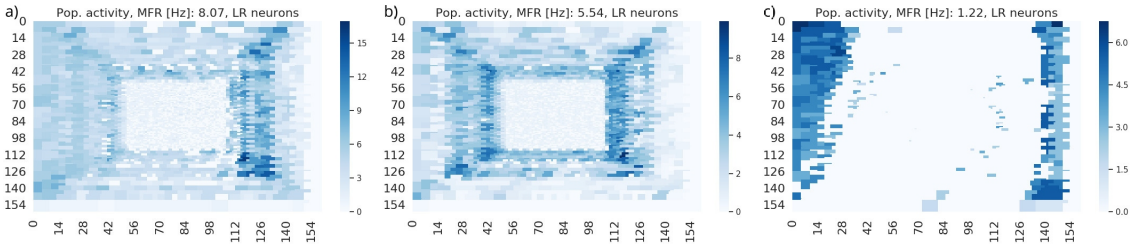


Figure B.1. Response from the population of LR sEMDs with the eccentric down-sampling mapped into the cartesian space with a camera resolution of 160x160 pixels. The color-code heatmap represents the MFR of each RF. Population response to real data for three symbolic speeds: slow (a), medium (b) and fast (c).

of increase for increasing stimulus velocities. A possible explanation for this is the ATIS cameras inherent noise around boarders of the moving bar causing false triggers at slow speeds. Furthermore, as the recordings were performed in an open space with many possible interference sources, thus the datasets are not ideal. However, the center of mass location of the RFs response still provides information about the stimulus velocity. Thus, showing that the model is suitable for a real-world application.