Top-down and bottom-up attentions in event data: a survey

Huyen Trang Nguyen[1], Laurent Sparrow[2,3][0000−0001−7388−9363], and Jean Martinet[1][0000−0001−8821−5556]

[1] Université Côte d'Azur, CNRS, I3S, France
{huyen-trang.nguyen, jean.martinet } @univ-cotedazur.fr
https://www.i3s.unice.fr/jmartinet

[2] Université de Lille, UMR 9193 - SCALab - Sciences Cognitives et Sciences Affectives
laurent.sparrow@univ-lille.fr
https://pro.univ-lille.fr/laurent-sparrow

Abstract. This survey explores the application of top-down and bottom-up attention mechanisms in event data. Event cameras, also known as dynamic vision sensors (DVS), capture changes in brightness at each pixel asynchronously, providing high temporal resolution, low latency, and efficient data handling, making them ideal for dynamic and real-time environments. This paper elaborates on the distinction between bottom-up attention, driven by sensory input and saliency, and top-down attention, guided by cognitive factors such as task relevance and prior knowledge. The application of these attention mechanisms to RGB data is discussed, providing a foundation for understanding their potential in event data. We highlight how these mechanisms can be integrated to enhance processing efficiency and accuracy in event vision systems. By leveraging insights from cognitive science and neuromorphic computing, we review existing models utilizing event data. These models aim to dynamically adjust attention based on task requirements, demonstrating their potential in applications such as robotics, autonomous driving, gesture recognition, and surveillance. The paper concludes by outlining the benefits of these advancements and the need for further research to refine these models for practical deployment.

Keywords: Cognitive attention · Top-down and bottom-up · Event data.

1 Introduction

In the field of computer vision, the quest for low-latency, energy-efficient vision systems has catalyzed the development of innovative models capable of processing vast amounts of visual data with high efficiency. Traditional frame-based cameras, despite their effectiveness, produce a fixed amount of data per frame, leading to redundancy and significant processing demands. Event cameras, also known as dynamic vision sensors (DVS), present an alternative by capturing only changes in brightness at each pixel, thereby reducing data redundancy and power consumption. These cameras offer high temporal resolution and low latency, making them particularly suitable for dynamic and real-time processing environments.

The concept of attention plays an important role in enhancing the performance of these vision systems. Attention mechanisms can be categorized into bottom-up attention, driven by the inherent saliency of sensory input, and top-down attention, guided by higher-level cognitive factors such as task relevance and prior knowledge. Bottom-up attention focuses on the inherent features of stimuli, while top-down attention integrates contextual information and cognitive biases to direct focus based on the task at hand. This paper aims to bridge the gap in current research by providing a review of selected approaches of top-down and bottom-up attention mechanisms as applied to event data.

The potential applications of these integrated attention mechanisms are vast, including robotics, autonomous driving, gesture recognition, and surveillance. Reducing computational load and improving response times, these systems can operate more efficiently in real-time scenarios, offering significant advancements in performance and energy efficiency. This survey aims to demonstrate the effectiveness of these integrated attention mechanisms in improving the performance of vision systems and outlines the benefits of these advancements and the need for further research to refine these models for practical deployment.

This survey paper is part of an interdisciplinary collaborative research project, *Neuromorphic Attention Models for Event Data* (NAMED[3]), that aims to develop computational models of cognitive attention that combine both event sensors (simulating peripheral vision) and RGB sensors (simulating central vision), with improved power efficiency in embedded electronics, mimicking the human visual system to selectively focus on regions of interest. By leveraging the efficiency of these mechanisms, the project seeks to optimize energy consumption in computer vision systems, benefiting applications such as robotics, autonomous vehicles, and wearable devices. This approach could lead to more sustainable and efficient systems in large-scale operations.

The structure of this paper is as follows: The first section 1 introduces the general discussion of visual attention mechanisms in event data. We continue with an overview of event cameras, event data, and datasets in section 2. The next section 3 examines the concepts of visual attention in computer vision, discussing regions of interest, eye gaze, and the interplay between bottom-up and top-down mechanisms. This is followed by a review of top-down and bottom-up attention mechanisms for RGB data in Section 4. The subsequent section 5 delves into the application of these mechanisms to event data. The paper concludes in section 6 with suggestions for future research directions.

[3] Project number ANR-23-CE45-0025-02, URL: https://wwww.i3s.unice.fr/named/

2 Event cameras and datasets

This section introduces event cameras and event data, and reviews over 20 existing event datasets categorised by the target computer vision task they address.

2.1 Event cameras and event data

Event cameras, also known as dynamic vision sensors, represent a significant shift from traditional frame-based cameras. Unlike frame-based cameras that capture image frames at fixed intervals, event cameras only record changes in brightness asynchronously and independently at each pixel [43]. These changes, termed *events*, are categorized as Positive or Negative (On or Off), corresponding to increases or decreases in brightness (or polarity). Figure 1 [43] illustrates the difference between data captured by a frame-based camera and an event camera. On the left is a screenshot taken from a video recording, showing an arm waving in front of a frame-based camera. In contrast, the image on the right is generated from event data, where only the arm is visible due to its motion causing changes in light. In this image, other objects such as the table and whiteboard are not captured by the event camera because they remain stable, resulting in no changes in brightness on these objects.

Event cameras possess several key features that distinguish them from traditional cameras. They capture data asynchronously, with each pixel independently detecting changes in light intensity and generating data only when movement or lighting changes occur, thereby reducing redundancy. Their high temporal resolution and low latency allow them to capture fast-moving objects with minimal motion blur. Event cameras also consume less power since they activate only when changes are detected. Additionally, they excel in high dynamic range environments, effectively handling scenes with rapid transitions between dark and bright conditions, where traditional cameras often struggle due to fixed exposure settings.

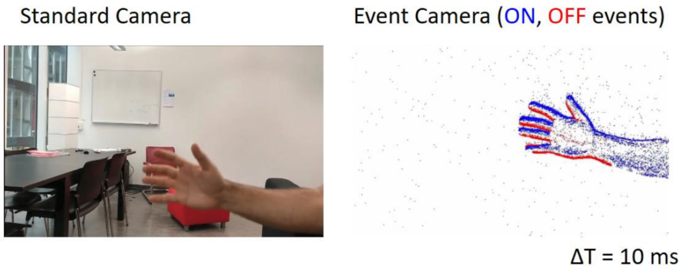

Fig. 1: Data captured by frame-based camera and event camera. Source: [43].

The data generated by event cameras is typically represented in the form of (x, y, t, p), where x and y denote coordinates, t represents the time stamp when light changes (*events*) occur, and p indicates polarity (positive or negative brightness change). This results in a stream of events rather than frame-based cameras, capturing only the dynamic parts of a scene. The data is inherently sparse, consisting only of changes in the scene, making it highly efficient but requiring different processing algorithms compared to typical video data. Each event is time-stamped, providing precise temporal information crucial for accurate measurements. The continuous data stream is well-suited for real-time processing, applicable in fields such as robotics and autonomous driving.

Event cameras present significant differences compared to traditional frame-based cameras, offering great potential for future research. In the following subsection, we discuss datasets for event data and their potential applications.

2.2 Event datasets

Existing event datasets, which have emerged in recent years, vary widely in structure and recording conditions, based on their specific needs and applications.

Event datasets for object recognition and 3D object perception include the Caltech-256 Dataset [20], which captures objects in motion with event cameras, and CIFAR10-DVS [27] and N-Caltech101 [38] which adapt CIFAR10 and Caltech101 images respectively for event cameras in recognizing categories. The Combined Dynamic Vision/RGB-D Dataset [53] provides scenarios for object recognition using synthetic, color, and depth sensors, while MVSEC [57] [58] and Event Camera Motion Segmentation Dataset offer extensive 3D object recognition and segmentation for various vehicles and objects indoor and outdoor. The N-CARS dataset supports specifically for cars classification [44]. Event-based Moving Object Detection and Tracking [36] includes event data for tracking moving objects. Slow-motion card symbol recognition is analysed by SLOW-POKER-DVS dataset. Besides object recognition, some event datasets supporting gesture recognition include the DHP19 [4] which presents a human pose dataset in event data. The DVSMOTION20 [1], ROSHAMBO17 [33] and SL-ANIMALS-DVS [50] [49] explore for human sign languages or hand moving.

For autonomous driving, DDD20 [19] and DSEC [14] [15] datasets offer extensive data on road scenes and vehicle movements in various brightness scenarios. The DET: A High-resolution DVS Dataset for Lane Extraction [6] focuses on lane extraction with labeled images. DND21: DeNoising Dynamic Vision Sensors Dataset [18] address action recognition and denoising for surveillance and driving use cases. Driving videos are further explored in the Driving Event Camera Dataset [42], including high speed driving datasets, and DVS09 [9] [30], which is utilized for simple outdoor driving in day-light condition. In addition, EDFLOW21 [32] focuses on driven-flow with event data. Pedestrian actions event data is analysed in Neuromorphic Vision Dataset for Pedestrian Detection, Action Recognition, and Fall Detection [35] and GEN1 Automotive Detection Dataset [8].

In summary, numerous event datasets offer significant potential for research across various applications and use cases. In the next section, we explore attention mechanisms for computer vision, focusing on both RGB and event data.

3 Attention for event data in computer vision

Since this survey discusses the concept of attention tailored to event data, it is important to clarify that within the scope of this paper, the term *attention* specifically pertains to the realms of cognitive attention. We highlight below the distinction between cognitive attention and attention mechanism of Transformer architecture for computer vision in general.

The attention mechanism [51] in the Transformer architecture operates originally in the domain of Natural Language Processing (NLP). At its core, the Transformer attention mechanism enables the model to capture long-range dependencies within sequences by weighing different parts of the input sequence differently, thereby focusing more on relevant information and disregarding noise. Additionally, recent advancements, such as the Vision Transformer (ViT) [10], have extended and adapted the attention mechanism from Transformer to computer vision by segmenting images into fixed-size patches, embedding them, and employing a Transformer encoder. This technique enhances image classification with attention-based processing.

Unlike attention for Transformers, cognitive attention involves higher-order cognitive processes beyond sensory perception, such as memory, decision-making, and problem-solving [48]. Cognitive attention mechanisms enables the brain to prioritize and efficiently process sensory information from the environment. It can manifest across various sensory modalities such as auditory and tactile perception. Tsotsos and colleagues define attention as “the process by which the brain controls and tunes information processing.” [46]. The human visual system only allows for high-resolution visual information to be encoded from the fovea (the central 2° of vision). Visual quality falls off rapidly and continuously from the center of gaze into a low-resolution visual surround. As a result, we constantly move our eyes (saccade) to redirect the fovea towards a new area where the visual information will be acquired when the eye is stable (fixation). Thus, due to the structure of our visual system, human vision depends on eye movements. Understanding the factors that guide these eye movements is therefore an important component of understanding how humans process visual information and has a wide range of applications in computer vision. A two-component framework for attentional deployment has emerged in the early 2000’, suggesting that the observer selectively directs attention to objects in a scene using both bottom-up, image-based cues, and top-down, task-dependent cues [22], as described below.

Visual attention is attracted by salient stimuli that *pop out* from their surroundings. Bottom-up, or stimulus-driven, selection is said to occur when attention captured by properties of the stimulus even if they are irrelevant to the current task. Some stimuli are intrinsically conspicuous or salient and spontaneously and involuntarily attract attention. Saliency, which is independent of the nature of the particular task, operates very rapidly. This suggests that saliency is computed in a pre-attentive manner across the entire visual field.

But attention can also be voluntarily directed to objects of current importance to the observer. Top-down, or goal-directed, selection is said to occur when the observer's knowledge or beliefs about the task determine what is selected in the visual field. Attention adapts the visual system to its dynamic needs. For example, when individuals are highly engaged in a problem-solving task (top down) or when they have to detect targets rapidly on a screen (bottom up), visual inputs are processed differently. When attempting to construct a general model of visual attention, one has to take into account the fact that these different strategies are more or less induced by the tasks performed by the participant.

In the real world, we constantly move our eyes to direct the high-resolution fovea towards points of interest in the environment. This situation is different from in-laboratory research tasks, where stimuli are displayed on a screen and the useful visual field is smaller. Such studies involve foveal rather than peripheral vision. The scientific literature is mainly focused on modelling bottom-up rather than top-down attention. But above all, it is much less common to find models that take both sources of information into account in a dynamic way [54].

In the following sub-sections, we delve into key concepts in attention for computer vision, such as top-down and bottom-up attention, Region of Interest, and eye gaze.

3.1 Bottom-up and top-down attentions

Bottom-up attention, often referred to as exogenous attention, is driven by sensory input and operates rapidly and involuntarily [25] [13]. One prominent concept in bottom-up attention is *saliency*, which denotes the degree to which a stimulus stands out from its surroundings [41]. Saliency-driven attention is guided by the conspicuousness of visual features such as color, orientation, and motion, with salient stimuli attracting attention spontaneously. Computational models, such as saliency maps, have been developed to simulate bottom-up attention, wherein saliency values are computed across the visual field, facilitating the detection of regions likely to capture attention [52]. These models provide insights into how visual stimuli compete for attentional resources based on their saliency, offering a framework for understanding early stages of visual processing.

In contrast to bottom-up attention, top-down attention is endogenous, driven by internal cognitive factors such as goals, expectations, and task relevance. This form of attention is characterized by its voluntary nature and its ability to prioritize specific stimuli based on cognitive goals and expectations [25] [13] [31]. Top-down attention can modulate sensory processing, biasing attention toward stimuli relevant to the observer's goals while filtering out irrelevant information. While bottom-up attention is primarily concerned with the saliency of sensory input, top-down attention operates at a higher cognitive level, integrating contextual information and cognitive biases to guide attentional selection [37] [7]. Modeling top-down attention poses challenges due to its complexity and its dependence on internal cognitive states, making it a subject of ongoing research in cognitive neuroscience and computational modeling.

3.2 Region of interest and eye gaze

Regions of Interests (RoI) and scan paths (sequence of points of gaze) provide complementary insights into attentional mechanisms and cognitive processing. The gaze is focused on a point on the scene or screen and RoI represent the zone of what is actually seen in a single glance in central vision. RoI selections, driven mainly by top-down attention and task-specific goals, offer a direct insight into cognitive processing and perceived interest in visual scenes [11]. In contrast, scanpaths reflect the interplay between bottom-up and top-down attention, providing temporal information about the observer's visual behavior [47]. While RoI selections provide spatially defined regions of interest based on cognitive factors, scanpaths offer continuous temporal information about the observer's attentional allocation during scene processing [11]. Despite differences in spatial coverage and temporal resolution, both approaches contribute to understanding attentional mechanisms and cognitive processes involved in visual perception.

4 Top-down and bottom-up attention for RGB data

In the domain of computer vision, much research has explored attention approaches for analyzing RGB data. In this section, we will briefly review top-down and bottom-up models for RGB data before shifting more deeply into those for event data in the section 5.

In 1998, Itti, Koch, and Niebur [23] introduce a computational model of visual attention that mimics the early stages of visual processing in primates. The model creates a saliency map by integrating various visual features such as intensity, color, and orientation at different spatial scales. These features are merged to emphasize the most visually prominent areas within a scene. The saliency map is then used to direct attention towards these regions, facilitating rapid analysis of the scene. The primary goal of this model is to replicate the way humans and primates quickly identify significant areas in their visual field without relying on prior knowledge or specific tasks. This model is classified as a bottom-up approach because it depends entirely on the characteristics of the visual stimuli to decide where to focus attention. In Itti et al.'s model, attention is given to areas that stand out due to distinct visual features, such as high contrast or unique colors, rather than any pre-determined expectations or goals. This approach is particularly effective for initial scene analysis, where quickly identifying salient features is essential.

The study by Martinet et al. (2009) [34] explores gaze analysis as a tool for evaluating the quality of visual media, focusing on a bottom-up visual attention approach. This method utilizes viewers' eye movements to determine how effectively visual media communicates its intended message. By tracking and analyzing where and how long viewers look at different parts of an image or video, the research identifies the elements that naturally attract attention. This approach depends on the visual characteristics of the media itself, such as color, texture, and movement, emphasizing a data-driven method that highlights the most eye-catching features and thus representing a bottom-up perspective in visual attention analysis.

Wang et al. (2010) [52] examines the link between bottom-up (or visual salience) and top-down (or visual importance) by two experiments. The first experiment involved participants rating the importance of hand-segmented objects in images. The second used eye-tracking to determine visual saliency based on where participants naturally looked in images without a task. The findings reveal a moderate correlation between importance maps and saliency maps. Saliency maps highlight shape and color, with early focus on human and animal faces. In contrast, importance maps prioritize categories such as human and animal faces, influenced by artistic meaning of the image. It also revealed a strong relationship between visual salience and visual importance in the first two seconds of viewing, indicating that bottom-up attention efficiently identifies primary subjects initially. This research highlights the complementary roles of bottom-up and top-down attention mechanisms.

The Deep Recurrent Attentive Writer (DRAW) network by Gregor et al. (2015)[16] for image generation features a unique spatial attention mechanism inspired by human eye foveation. The architecture of DRAW model contains encoder and decoder, both are recurrent networks, which iteratively build images through sequential modifications. The network uses dynamically updated attention, focusing selectively on input and output regions with 2D Gaussian filters to create image patches with varying locations and zoom levels. This bottom-up attention allows DRAW to adapt to different object scales and positions based on database, enhancing its performance in complex visual tasks.

Zheng et al. (2015) [56] developed Fixation NADE model, which uses visual attention by directing the recognition process through a sequence of task-specific fixations. Unlike traditional models that densely extract features over an entire image, Fixation NADE simulates human visual attention by learning both what features to extract and where to extract them, using a fixation policy. This approach combines both bottom-up and top-down attention mechanisms, where bottom-up attention is influenced by the data observed at each fixation point, and top-down attention is guided by task-specific objectives (gender classification and expression classification). Key contributions of this model include its autoregressive architecture for fixation-based recognition and its improved performance compared to earlier models like Fixation RBM [12].

CLIP (Contrastive Language-Image Pre-Training) model by Radford et al. (2021) [40] aims to develop visual representations using natural language supervision and employs a contrastive learning method to align images with their corresponding text descriptions. The goal of CLIP is to harness the rich information in natural language to enable zero-shot transfer learning, allowing the model to adapt to a variety of tasks without additional labeled data. The architecture includes an image encoder and a text encoder that work together to identify the correct image-text pairs. CLIP excels in multiple computer vision benchmarks, demonstrating strong performance in tasks such as action recognition, image description, or fine-grained object classification etc. The model leverages top-down attention by using textual descriptions to guide the focus on relevant image features. This innovative approach highlights the potential of natural language supervision in developing versatile and transferable visual models.

5 Top-down and bottom-up attention for event data

In contrast to RGB data, research of attention focusing on event data has been less prevalent, primarily due to its unique characteristics that demand specialized analysis. While there exists some studies of bottom-up, it is noteworthy that research on top-down models for event data remains relatively novel, with only a limited number of existing work. This could be due to the complexity of top-down approaches where the model depends significantly on specific predefined-tasks. However, top-down models remains a promising avenue for future research.

Li's 2019 paper [28] combines event data with frame-based images to enhance object detection for autonomous vehicles. The model employs two distinct processing streams: one utilizes convolutional neural networks (CNNs) for frame-based data, and the other leverages spiking neural networks (SNNs) for event data, which produce visual attention maps. By combining the high temporal resolution and dynamic range of event cameras with the detailed spatial information from traditional cameras, this approach aims to improve detection accuracy, particularly in scenarios with rapid motion and varying lighting conditions. The fusion of outputs from both streams is managed using Dempster-Shafer theory [26], resulting in improved performance as demonstrated on the DDD17 dataset [2]. The event vision system in this paper operates using a bottom-up approach. This method captures data only when there are changes in the visual field, which naturally highlights regions with motion and activity. This attention mechanism enhances the model's ability to detect vehicles effectively in challenging environments, highlighting the benefits of combining neuromorphic and conventional vision technologies in autonomous driving.

Iacono et al. (2019) [21] focuses on modifying a proto-object attention model to function with neuromorphic event cameras, aiming to enhance the iCub humanoid robot's visual processing capabilities. Utilizing a bottom-up attention mechanism, the model processes the event data through three layers: center-surround filtering, border ownership cells, and grouping cells. These layers decompose the visual scene into proto-objects, with saliency determined by changes in contrast and edges. The approach is validated with both static and moving objects, showing that higher object speeds lead to increased event counts, which enhances saliency. This model offers an efficient, low-latency method for robotic vision in dynamic settings, ensuring that the robot can effectively focus on relevant stimuli in its environment.

Yao et al.'s 2021 research [55] presents an innovative method for processing spatio-temporal event streams using Spiking Neural Networks (SNNs). The main goal of the study is to improve the accuracy and efficiency of SNNs by incorporating a temporal-wise attention mechanism. This model, named TA-SNN (Temporal-wise Attention Spiking Neural Networks), selectively filters out non-essential event streams (or denoising) during the inference phase based on their importance, which is determined during the training phase. This optimizes computational resource usage. The model operates primarily with a bottom-up attention mechanism, as it processes event data by dynamically focusing on significant temporal segments determined by the data itself. The TA-SNN model achieved state-of-the-art results in various classification tasks, such as gesture recognition, image classification, and spoken digit recognition, demonstrating its effectiveness in handling sparse and uneven event streams.

In their paper in 2021, Liang et al. [29] introduces an innovative approach for event-based object detection that employs a lightweight spatial attention mechanism. The primary goal is to improve the accuracy and efficiency of detecting objects in event data, especially by minimizing noise and enhancing multi-scale feature maps through the integration of shallow features. The core method involves encoding the event data into maps using techniques like Surface of Active Events (SAE) and Histogram of Intensities (HIS), followed by the application of a Canny edge detector to highlight relevant features. A key innovation is the Spatial Attention Module (SAM), which integrates multi-scale spatial features with the object detection framework. This module focuses on areas with significant event activity, enhancing the detection of moving objects by concentrating computational resources on RoI. The model utilizes a bottom-up approach, starting with raw event data and applying spatial attention to highlight significant features. By leveraging this lightweight attention mechanism, the system effectively filters out noise and improves detection accuracy, ensuring better performance in real-time automotive scenarios where rapid response and accuracy are crucial.

In the realm of autonomous driving, attention using event data is crucial for improving the system's ability to handle dynamic environments and ensure safety. Sun's 2022 research [45] introduces the Event Fusion Network (EFNet) designed to mitigate motion blur in images captured by conventional frame-based cameras. EFNet employs a symmetric cumulative event representation and integrates an Event-Image Cross-modal Attention (EICA) module to effectively blend data from both event cameras and frame-based cameras. The model operates in a two-stage architecture inspired by the event-based deblurring physical model [39], using bottom-up attention to highlight pertinent features from the event stream for accurate deblurring. This bottom-up method is driven by the event cameras' data, which record intensity changes with high temporal resolution, enabling the model to focus on crucial motion information for deblurring. Additionally, the paper introduces the REBlur dataset, featuring real-world event streams paired with corresponding blurry and sharp images, to assess EFNet's performance in challenging scenarios. This study marks a significant advancement in motion deblurring by effectively utilizing event data, following bottom-up approach.

Gesture recognition is another application where attention using event data proves to be highly effective. Gruel's 2022 study [17] demonstrate the advantages of using neuromorphic attention for event data for recognizing hand and arm movements. Gruel's method involves a spiking neural network that dynamically adapts to incoming event streams, focusing on regions of high activity that correspond to meaningful areas while filtering out irrelevant background noise. The regions with the highest density of event data are defined as the focal points of attention (RoI) within the respective model, enabling the system to prioritize areas with significant activity while discarding other unrelevant locations.

In the context of object recognition from a static camera, attention using event data offers certain advantages, especially in scenarios involving dynamic environments and varying lighting conditions. In Bulzomi's 2023 study [3], pedestrian detection is explored using a static camera setup equipped with an event-based sensor. The dataset includes various recordings where event data is processed to filter out background noise and focus on the movement of pedestrians. This method leverages the asynchronous nature of event data to detect changes in the scene, allowing the system to efficiently isolate and recognize moving objects, such as pedestrians, against a static background. The definition of attention in Bulzomi's study aligns with Gruel's work [17], where RoI is determined as the position that contains the highest density of events, ensuring that the system concentrates on the most pertinent data.

Cancini's 2019 research [5] presents a novel approach for object recognition utilizing event cameras, focusing on enhancing processing efficiency in dynamic environments. The authors propose two models: the first model tracks event activity to identify RoI using a peak detection algorithm, implementing a top-down approach to focus computational efforts on the most active areas. The second model adapts the DRAW-based neural architecture [16] to handle event data effectively. These RoIs are then processed by a Phased LSTM recognition network, which combines spatial and temporal information for accurate object classification. The significant contribution of this research lies in the adaptation of attention mechanisms specifically for event cameras, demonstrating improved handling of translation and scale variations compared to conventional models.

Kong et al. (2024) [24] introduce a model for event-based semantic segmentation, leveraging the pre-trained knowledge from image and text domains using CLIP [40]. The objective is to address the challenges of dense annotations and scalability in event-based vision tasks by transferring semantically rich CLIP knowledge to event streams. The model employs a top-down approach, using text prompts to guide the attention to relevant regions of the event data, facilitating zero-shot segmentation. This top-down mechanism effectively aligns event features with textual descriptions, allowing for semantic coherence in the segmentation process.

Here again, most attention approaches for event data focus on bottom-up mechanisms, with few incorporating top-down attention guided by task relevance. Amongst the state-of-the-art papers described in this section and the previous section, we notice that the majority of work mainly deal with bottom-up attention, possibly because it is harder and not straightforward to model top-down attention.

6 Conclusion

Event cameras, with their useful ability to capture only changes in brightness, offer high temporal resolution, low latency, and efficient data handling, making them ideal for dynamic and real-time environments. This survey highlights the distinct roles of bottom-up attention, driven by sensory input and saliency, and top-down attention, guided by cognitive factors such as task relevance and prior knowledge, in enhancing the performance of vision systems. The integration of top-down and bottom-up attention mechanisms in event data processing has the potential to increased efficiency and accuracy in embedded computer vision applications.

This review of existing research shows that while bottom-up attention mechanisms are well-explored and implemented, there is a significant need to develop and integrate top-down attention approaches. These approaches leverage contextual information and task-specific goals to prioritize relevant stimuli, enhancing the efficiency and accuracy. By combining both mechanisms, a more effective processing framework can be achieved, mimicking the human visual system's ability to selectively focus on important aspects of the visual field.

The potential applications of these integrated attention mechanisms are extensive, including robotics, autonomous driving, gesture recognition, and surveillance. By reducing computational load and improving response times, these systems can operate more efficiently in real-time scenarios, offering substantial advancements in performance and energy efficiency. Future research should continue to explore and refine these models, focusing on developing hybrid neuromorphic and deep learning approaches that can dynamically adjust attention based on the context and requirements of the task at hand. The exploration of top-down and bottom-up attention mechanisms in event data is promising for advancing the capabilities of modern vision systems. This survey provides a foundation for future research, highlighting the importance of integrating these mechanisms to create more efficient, and accurate vision systems.

Acknowledgments

This work was supported by the project NAMED (ANR-23-CE45-0025-01) of the French National Research Agency (ANR). This work was supported by the French government through the France 2030 investment plan managed by the National Research Agency (ANR), as part of the Initiative of Excellence Université Côte d'Azur under reference number ANR- 15-IDEX-01.

References

1. Almatrafi, M., Baldwin, R., Aizawa, K., Hirakawa, K.: Distance surface for event-based optical flow. IEEE transactions on pattern analysis and machine intelligence **42**(7), 1547–1556 (2020)

2. Binas, J., Neil, D., Liu, S.C., Delbruck, T.: Ddd17: End-to-end davis driving dataset. arXiv preprint arXiv:1711.01458 (2017)
3. Bulzomi, H., Gruel, A., Martinet, J., Fujita, T., Nakano, Y., Bendahan, R.: Object detection for embedded systems using tiny spiking neural networks: Filtering noise through visual attention. In: 2023 18th International Conference on Machine Vision and Applications (MVA). pp. 1–5. IEEE (2023)
4. Calabrese, E., Taverni, G., Awai Easthope, C., Skriabine, S., Corradi, F., Longinotti, L., Eng, K., Delbruck, T.: Dhp19: Dynamic vision sensor 3d human pose dataset. In: CVPR workshops. pp. 0–0 (2019)
5. Cannici, M., Ciccone, M., Romanoni, A., Matteucci, M.: Attention mechanisms for object recognition with event-based cameras. In: 2019 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2019)
6. Cheng, W., Luo, H., Yang, W., Yu, L., Chen, S., Li, W.: Det: A high-resolution dvs dataset for lane extraction. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition workshops (2019)
7. Connor, C.E., Egeth, H.E., Yantis, S.: Visual attention: bottom-up versus top-down. Current biology **14**(19) (2004)
8. De Tournemire, P., Nitti, D., Perot, E., Migliore, D., Sironi, A.: A large scale event-based detection dataset for automotive. arXiv preprint arXiv:2001.08499 (2020)
9. Delbruck, T., et al.: Frame-free dynamic digital vision. In: Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society. vol. 1, pp. 21–26. Citeseer (2008)
10. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
11. Engelke, U., Le Callet, P.: Perceived interest and overt visual attention in natural images. Signal Processing: Image Communication **39** (2015)
12. de Freitas, N.: Learning where to attend with deep architectures for image tracking. Neural Computation **24**(8) (2012)
13. Frintrop, S., Rome, E., Christensen, H.I.: Computational visual attention systems and their cognitive foundations: A survey. ACM Transactions on Applied Perception (TAP) **7**(1), 1–39 (2010)
14. Gehrig, M., Aarents, W., Gehrig, D., Scaramuzza, D.: Dsec: A stereo event camera dataset for driving scenarios. IEEE Robotics and Automation Letters **6**(3) (2021)
15. Gehrig, M., Millhäusler, M., Gehrig, D., Scaramuzza, D.: E-raft: Dense optical flow from event cameras. In: 2021 International Conference on 3D Vision (3DV). IEEE (2021)
16. Gregor, K., Danihelka, I., Graves, A., Rezende, D., Wierstra, D.: Draw: A recurrent neural network for image generation. In: ICML (2015)
17. Gruel, A., Vitale, A., Martinet, J., Magno, M.: Neuromorphic event-based spatio-temporal attention using adaptive mechanisms. In: 2022 IEEE 4th international conference on artificial intelligence circuits and systems (AICAS). IEEE (2022)
18. Guo, S., Delbruck, T.: Low cost and latency event camera background activity denoising. IEEE Transactions on Pattern Analysis and Machine Intelligence (2022)
19. Hu, Y., Binas, J., Neil, D., Liu, S.C., Delbruck, T.: Ddd20 end-to-end event camera driving dataset: Fusing frames and events with deep learning for improved steering prediction. In: 2020 IEEE 23rd International Conference on Intelligent Transportation Systems (ITSC). pp. 1–6. IEEE (2020)
20. Hu, Y., Liu, H., Pfeiffer, M., Delbruck, T.: Dvs benchmark datasets for object tracking, action recognition, and object recognition. Frontiers in neuroscience (2016)

21. Iacono, M., D'Angelo, G., Glover, A., Tikhanoff, V., Niebur, E., Bartolozzi, C.: Proto-object based saliency for event-driven cameras. In: IROS. IEEE (2019)
22. Itti, L., Koch, C.: Computational modelling of visual attention. Nature reviews neuroscience **2**(3) (2001)
23. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. IEEE TPAMI **20**(11), 1254–1259 (1998)
24. Kong, L., Liu, Y., Ng, L.X., Cottereau, B.R., Ooi, W.T.: Openess: Event-based semantic scene understanding with open vocabularies. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2024)
25. Le Callet, P., Niebur, E.: Visual attention and applications in multimedia technologies. Proceedings of the IEEE **101**(9), 2058–2067 (2013)
26. Lee, H., Kwon, H., Robinson, R.M., Nothwang, W.D., Marathe, A.M.: Dynamic belief fusion for object detection. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV). IEEE (2016)
27. Li, H., Liu, H., Ji, X., Li, G., Shi, L.: Cifar10-dvs: an event-stream dataset for object classification. Frontiers in neuroscience **11**, 244131 (2017)
28. Li, J., Dong, S., Yu, Z., Tian, Y., Huang, T.: Event-based vision enhanced: A joint detection framework in autonomous driving. In: 2019 ieee international conference on multimedia and expo (icme). IEEE (2019)
29. Liang, Z., Chen, G., Li, Z., Liu, P., Knoll, A.: Event-based object detection with lightweight spatial attention mechanism. In: 2021 6th IEEE International Conference on Advanced Robotics and Mechatronics (ICARM). IEEE (2021)
30. Lichtsteiner, P., Posch, C., Delbruck, T.: A 128x128 120db 15mus latency asynchronous temporal contrast vision sensor. IEEE J. of solid-state circuits (2008)
31. Lindsay, G.W.: Attention in psychology, neuroscience, and machine learning. Frontiers in computational neuroscience **14** (2020)
32. Liu, M., Delbruck, T.: Edflow: Event driven optical flow camera with keypoint detection and adaptive block matching. IEEE Transactions on Circuits and Systems for Video Technology **32**(9), 5776–5789 (2022)
33. Lungu, I.A., Corradi, F., Delbrück, T.: Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo. In: 2017 IEEE International Symposium on Circuits and Systems (ISCAS). IEEE (2017)
34. Martinet, J., Lablack, A., Lew, S., Djeraba, C.: Gaze based quality assessment of visual media understanding. In: 1st International Workshop on Computer Vision and Its Application to Image Media Processing (WCVIM) in conjunction with the 3rd Pacific-Rim Symposium on Image and Video Technology (PSIVT), Tokyo-Japan (2009)
35. Miao, S., Chen, G., Ning, X., Zi, Y., Ren, K., Bing, Z., Knoll, A.: Neuromorphic vision datasets for pedestrian detection, action recognition, and fall detection. Frontiers in neurorobotics **13** (2019)
36. Mitrokhin, A., Fermüller, C., Parameshwara, C., Aloimonos, Y.: Event-based moving object detection and tracking. In: IROS. pp. 1–9. IEEE (2018)
37. Noudoost, B., Chang, M.H., Steinmetz, N.A., Moore, T.: Top-down control of visual attention. Current opinion in neurobiology **20**(2) (2010)
38. Orchard, G., Jayawant, A., Cohen, G.K., Thakor, N.: Converting static image datasets to spiking neuromorphic datasets using saccades. Frontiers in neuroscience **9** (2015)
39. Pan, L., Scheerlinck, C., Yu, X., Hartley, R., Liu, M., Dai, Y.: Bringing a blurry frame alive at high frame-rate with an event camera. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (2019)

40. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: ICML (2021)
41. Rai, Y., Le Callet, P., Cheung, G.: Quantifying the relation between perceived interest and visual salience during free viewing using trellis based optimization. In: 2016 IEEE 12th Image, Video, and Multidimensional Signal Processing Workshop (IVMSP). IEEE (2016)
42. Rebecq, H., Ranftl, R., Koltun, V., Scaramuzza, D.: High speed and high dynamic range video with an event camera. IEEE transactions on pattern analysis and machine intelligence **43**(6), 1964–1980 (2019)
43. Scaramuzza, D.: Tutorial on event-based cameras. In: IROS 2015: Proc. of the 2nd Workshop on Alternative Sensing for Robot Perception (2015)
44. Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., Benosman, R.: Hats: Histograms of averaged time surfaces for robust event-based object classification. In: CVPR (2018)
45. Sun, L., Sakaridis, C., Liang, J., Jiang, Q., Yang, K., Sun, P., Ye, Y., Wang, K., Gool, L.V.: Event-based fusion for motion deblurring with cross-modal attention. In: European conference on computer vision. Springer (2022)
46. Tsotsos, J.K.: A computational perspective on visual attention. MIT Press (2021)
47. Tu, D., Min, X., Duan, H., Guo, G., Zhai, G., Shen, W.: End-to-end human-gaze-target detection with transformers. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE (2022)
48. Vaishnav, M.: Exploring the role of (self-) attention in cognitive and computer vision architecture. Ph.D. thesis, Université Paul Sabatier-Toulouse III (2023)
49. Vasudevan, A., Negri, P., Di Ielsi, C., Linares-Barranco, B., Serrano-Gotarredona, T.: Sl-animals-dvs: event-driven sign language animals dataset. Pattern Analysis and Applications pp. 1–16 (2022)
50. Vasudevan, A., Negri, P., Linares-Barranco, B., Serrano-Gotarredona, T.: Introduction and analysis of an event-based sign language dataset. In: International Conference on Automatic Face and Gesture Recognition (2020)
51. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. Advances in neural information processing systems **30** (2017)
52. Wang, J., Chandler, D.M., Le Callet, P.: Quantifying the relationship between visual salience and visual importance. In: Human vision and electronic imaging XV. vol. 7527. SPIE (2010)
53. Weikersdorfer, D., Adrian, D.B., Cremers, D., Conradt, J.: Event-based 3d slam with a depth-augmented dynamic vision sensor. In: 2014 IEEE international conference on robotics and automation (ICRA). pp. 359–364. IEEE (2014)
54. Yang, Z., Mondal, S., Ahn, S., Xue, R., et al.: Unifying top-down and bottom-up scanpath prediction using transformers. In: CVPR (2024)
55. Yao, M., Gao, H., Zhao, G., Wang, D., Lin, Y., Yang, Z., Li, G.: Temporal-wise attention spiking neural networks for event streams classification. In: CVPR (2021)
56. Zheng, Y., Zemel, R.S., Zhang, Y.J., Larochelle, H.: A neural autoregressive approach to attention-based recognition. IJCV **113** (2015)
57. Zhu, A.Z., Thakur, D., Özaslan, T., Pfrommer, B., Kumar, V., Daniilidis, K.: The multivehicle stereo event camera dataset: An event camera dataset for 3d perception. IEEE Robotics and Automation Letters **3**(3), 2032–2039 (2018)
58. Zhu, A.Z., Yuan, L., Chaney, K., Daniilidis, K.: Ev-flownet: Self-supervised optical flow estimation for event-based cameras. arXiv preprint arXiv:1802.06898 (2018)