An Object Tracking Using a Neuromorphic System Based on Standard RGB Cameras

E.B. Gouveia[1], L.M. Vasconcelos[1], E.L.S. Gouveia[1], V.T. Costa[1], A. Nakagawa-Silva[1] and A.B. Soares[1]

[1] Federal University of Uberlandia, Biomedical Engineering Laboratory - BIOLAB, Uberlandia, Brazil

Abstract— **Event-based cameras are devices that can be the key to solving various robotics challenges. However, unlike the Computer Vision field, research in Neuromorphic Vision does not have enough data for algorithms to be tested, evaluated, and compared to guarantee progress in the development of robust and competitive solutions. In this way, we propose the development of a framework that converts information recorded by standard RGB cameras into neuromorphic information. Using this framework, we create neuromorphic recordings from videos used in Computer Vision datasets to test and evaluate our tracking algorithm in neuromorphic recordings. We reached an average accuracy of 95.38% in tracking the information of the five videos selected in this work.**

Keywords— **event-based, cameras, tracking, neuromorphic.**

I. Introduction

Event-based systems have become an attractive field for the development of new works related to the challenges of robotics and new computational approaches [1] [2] [3]. Neuromorphic devices consist of bioinspired approaches to the development of technology.

One of the neuromorphic devices that has gained space in recent years has been event-based cameras [4] [5] [6]. These vision sensors work similarly to the retina and have advantages over standard cameras. Each pixel works independently and asynchronously transmitting binary events, allowing a low dimensionality of data, high temporal resolution, low latency, and low energy consumption [7]. These characteristics guarantee a wide advantage of neuromorphic sensors over standard vision sensors in tasks such as motion estimation [8] [9] and tracking [10].

Being a new field of study, it is necessary to standardize and democratize neuromorphic information so that the work in the young field of Neuromorphic Vision reaches sufficient development so that algorithms can be compared with each other [11].

An approach for converting information recorded by standard cameras into neuromorphic information was described by [12]. This approach allows the algorithms developed for Neuromorphic Vision to be used in Computer Vision benchmark datasets and take advantage of their ground truth information.

In this work, we developed a framework for converting information recorded by standard RGB cameras into neuromorphic information. We convert videos present in Computer Vision datasets, with ground truth information for validation of algorithms, in videos based on Neuromorphic Vision. We apply a tracking and segmentation algorithm to neuromorphic information and evaluate the performance of the algorithm.

II. Materials and Methods

A. Standard RGB to neuromorphic

The conversion of the information recorded by a traditional camera to a neuromorphic model was done using a framework developed in python. Our framework can be used, in real time, to convert recordings using any traditional camera, or even to convert pre-recorded videos.

In neuromorphic hardware, each event is generated asynchronously through a comparator circuit that receives the luminosity information from a photodiode and, when detecting a significant variation in luminosity in relation to the basal luminosity of the scene, triggers an event.

Each event captured by these neuromorphic vision sensors carries three basic information: a) the luminance variation (positive or negative) b) the temporal moment at which the event occurred and c) the location at S(x, y) at which contrast variation has occurred [4].

In our framework (Fig. 1), the conversion of traditional information to neuromorphic is done through the analysis of the luminosity variation at different instants (Equation 1), however, the temporal resolution of the neuromorphic recordings is linked to the transmission speed in Frames per Second (FPS) of the camera used to make these traditional recordings.

The process of converting traditional information into neuromorphic was developed in a parameterized way, where it is possible to determine the resolution of the neuromorphic recording and the sensitivity to the generation of events, in other words, it is possible to determine the minimum luminance variation required to generate an event.

$$g(t) = \begin{cases} K, & \text{if } f(t) - f(t-1) \geq \beta x K \\ 0, & \text{if } f(t) - f(t-1) \leq -\beta x K \\ \frac{K}{2}, & \text{otherwise} \end{cases} \quad (1)$$

where K is the maximum possible luminosity value according to the bit depth of the recording, beta is the sensitivity parameter value, which can vary from zero to one, f(t) is the luminance value at the current instant, f(t-1) is the luminosity value in the previous instant and g(t) is the neuromorphic frame value in the current instant.

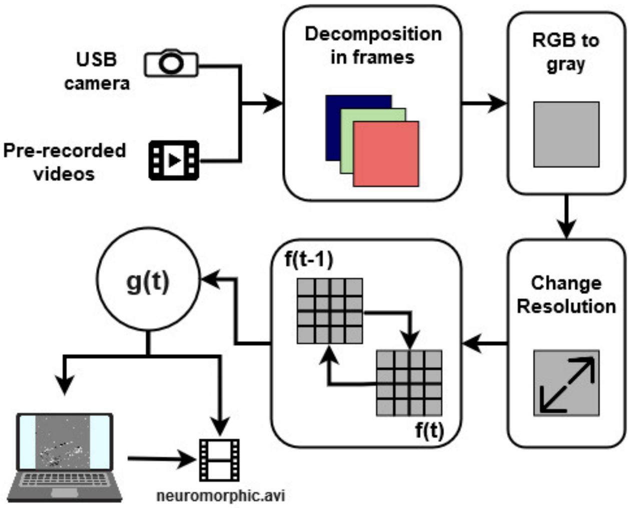

Fig. 1: Diagram of our framework.

B. Dataset

The recordings used in this work were chosen from datasets used in traditional Computational Vision for tracking and segmentation [13].

Given the core of neuromorphic vision sensors, one of the great advantages of these devices lies in minimizing the redundant information of static scenes by generating events only when there is variation in luminosity. Therefore, we chose to use recordings that have contexts similar to the context in which the properties of a neuromorphic vision sensor were more appropriate, being them a) scenes with static background information, b) scenes with relative movement between the target and the camera and c) scenes where it is of interest screening objects.

Therefore, five videos were selected from different datasets to submit them to the process of generating neuromophical recordings: ”Vid A Ball” [14], ”Walking” [15], the first video from the Hallway dataset (Hallway-1) [16], lemming [17] and Venice-1 [18]. Each video used in this work has the ground truth of the ideal bounding boxes in the tracking and segmentation process. The ground truth of these videos was used to evaluate the characteristics of the algorithm developed in this work, as it will be discussed in the following sections.

C. Watershed Segmentation Method

To perform the segmentation of objects, the Watershed method [19] [20] [21] was used, which consists of a morphological and region-based segmentation process. An image analysis approach is used considering a surface topography based on shades of gray, where the gray value of the pixel represents the elevation of the landscape.

In this way, several points of local minimums are created where all other pixels that present values comprised in a descending gradient that end in a point of local minimum are grouped and form a region (a catchment basin) that are separated by watersheds [22].

However, by assigning a basin or watershed label to each pixel, noisy images will result in several small regions, which generates an over-segmentation phenomenon. The Watershed method has been used widely and has shown good results for segmentation processes in Computer Vision [23] [24] [25] [26] and has also been used for the segmentation of medical images [22] [27] [28] [29].

Neuromorphic images have binary information and the grouping of this generic information suggests more complex information than an isolated event. Therefore, morphological approaches can be an interesting methodology to be applied in neuromorphic recordings. However, noise present in neuromorphic recordings causes a constant over-segmentation. To solve this problem, we a) run a threshold-filter for small regions, b) create a rectangular bounding box covering the contour of different regions and merge the bounding boxes that intersect each other, and c) merge bounding boxes with nearby centroids, given a distance threshold.

D. Quantitative analysis

Accuracy: To calculate the accuracy of our tracking system we performed an analysis of the ratio between the number of frames that have an Intersection over Union (IoU) between the bounding box of the ground truth and the bounding box estimated by our algorithm, and the total frames of the video.

Tracking evaluation: To assess the ability of our algorithm to track an object, we analyzed the distance between the centroid of the bounding box on the ground truth and the centroid of the bounding box estimated by our algorithm.

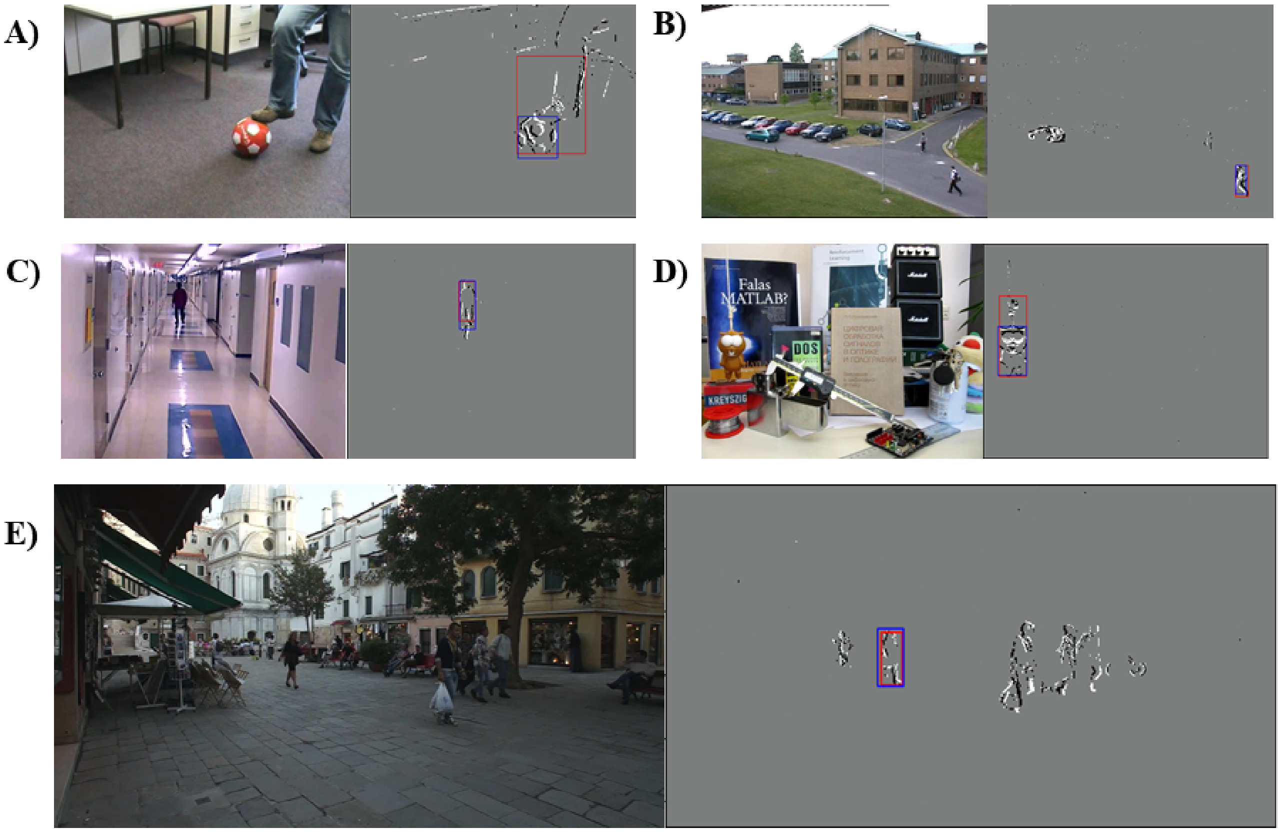

Fig. 2: Result of the process of converting images recorded by traditional sensors into neuromorphic recordings. a) Vid a Ball, b) Walking, c) Hallway 1, d) lemming and e) Venice 1. To the left it is shown the traditional recording, to the right the neuromorphic frame. The blue bounding boxes are the ground truth and the red ones the result of our segmentation.

However, as we use several videos from different datasets, there is a disparity in the distance metric when it comes to recordings with different resolutions. With this, we perform the normalization of the distance between the centroids using the diagonal value of each video, in other words, we measure the tracking capacity of our system through the percent analysis of the distance between the centroids in relation to the largest diagonal of the frame.

Segmentation evaluation: To analyze our system's ability to segment the scene, we performed an IoU check throughout the video. The quantization was done by distributing the frequency of the percentage of IoUs. In other words, we analyzed the percentage of frames that have IoU between 0 and 10%, 10 and 20% and so forth.

III. Results and Discussion

The recordings made by traditional RGB cameras and converted to a neuromorphic format can be seen in Fig 2. On the left side, ir is shown a frame of the original recording, while on the right side is the result of the conversion process using our framework. In Fig 2, the neuromorphic frames also have the marking of both bounding boxes, in blue (ground truth) and in red (our algorithm). Each pair in Fig. 1 shows a frame from the videos used in this research: a) Vid a Ball, b) Walking, c) Hallway 1, d) lemming and e) Venice 1.

Since the research on Neuromorphic Vision sensors began, some papers have highlighted the importance of creating and consolidating neuromorphic datasets [11] [30]. Good neuromorphic datasets are necessary so that algorithms in Neuromorphic Vision can be performed and compared. This work in Biomedical Engineering follows similar steps as those in the mature field of Computer Vision [12] [31] [32] [33], which through the global collaboration to generate large datasets with labeled information made available to the community.

There are three types of neuromorphic datasets in the literature. True-neuromorphic datasets, which consist of real scenes recorded with event-based sensors [34] [35], Computer Vision datasets that are converted to Neuromorphic Vision datasets through a process that uses an event-based sensor and a monitor which presents the dataset information to the event-based sensor and in this way the recordings are "translated" [30] [36] and there those generated by fast software approach that can, later, be implemented in hardware (such as FPGAs), such as those presented in this work and other researches [37] [38], using different manipulations to generate pseudo-neuromorphic information from traditional RGB scenes.

We applied a tracking algorithm to the pseudo-neuromorphic recordings, and performed the analysis of its tracking capacity (Table 1), localization (Table 2, Fig 3) and segmentation (Fig 4). Table 1 shows the accuracy of our tracking system, through the analysis of the percentage of detections that have an intersection with the ground truth bounding box.

Table 1: Accuracy of tracking

Video	Accuracy
Vid A ball	97.34 %
Walking	95.45 %
Hallway 1	90.72 %
Lemming	99.33 %
Venice 1	94.03 %

The results for the accuracy capacity summarize the algorithm's ability to segment and track the object in the scene. However, to complement the performance estimate of the algorithm in the pseudo-neuromorphic recordings, we made an analysis of the distance between the centroid from the ground truth to the centroid of the detection estimated by our algorithm based on the Watershed method Table 2) shows the values of the mean distance between the centroids, with the standard deviation (σ), in percentage relative to the diagonal of the frames. A diagonal normalization process is necessary for the comparison between the results, since the resolution of each recording is different.

Table 2: Mean distance between centroids, as a percentage of the maximum possible distance in each frame

Video	Mean distance ± σ
Vid A ball	7.67 ± 7.97
Walking	0.44 ± 0.38
Hallway 1	1.56 ± 1.18
Lemming	4.63 ± 3.04
Venice 1	4.17 ± 5.09

When analyzing Table 2, it can be seen that there is a great variation in the average distance between the centroids, but in Fig 3 shows that, for more than 90% of the frames of the recording, the distances are less than 10% of the maximum distance of the frame, except for the Vid a Ball recording, which has a lower distance performance.

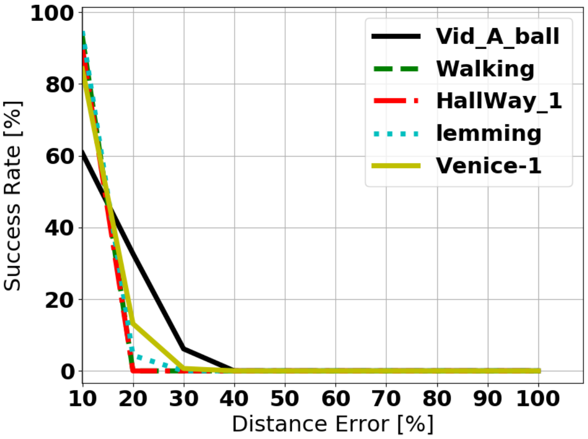

Fig. 3: Distribution of the distance between the bounding box of the ground truth and the controid of the bounding boxes estimated by our algorithm

The low relative performance of ”Vid a Ball” compared with the other videos is a consequence of the region-merge process used to solve the over-segmentation problem. In Fig 2 - A it can be seen that some times, there is an overestimation of the bounding box, which covers the ball and part of the leg of a subject who interacts with the ball (similar process happens in the lemming video, as observed in Fig 2 - D)

The performance of the segmentation process provided by our algorithm is shown in Fig 4. It is possible to see that the value of the IoU is concentrated around 40-90% in Hallway 1, lemming and Walking recordings, while for the Venice 1 and Vid a Ball recordings there is a concentration of IoU of less than 10%.

This overestimation of the space occupied by an object in the scene is due to the process of merging regions (as discussed earlier) and because the neuromorphic information does not present a gradual contrast variation, containing only binary information about the variation of light intensity, which does not allow for a significant region depth distinction. Thus, the Watershed method could not provide several regions with local minimums at different heights [22], which caused an over-segmentation process that was eventually corrected by our model to group nearby regions.

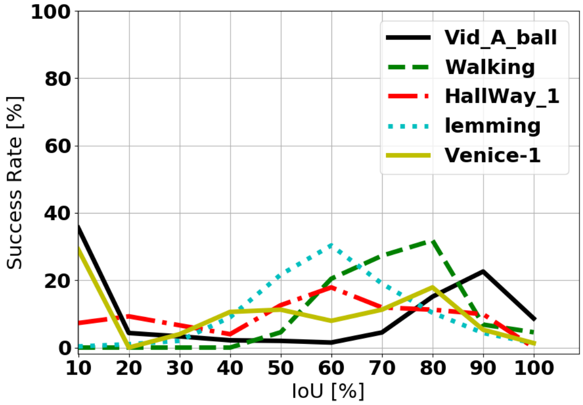

Fig. 4: IoU percentage between ground truth bounding boxes and our algorithm bounding boxes.

IV. Conclusion

In this work, we present a tracking system applied to pseudo-neuromorphic recordings converted from recordings made by traditional RGB cameras. Our system has been validated by submitting the recordings of conventional datasets to our framework, applying a tracking algorithm to the pseudo-neuromorphic information and evaluating the performance of our system with regard to segmenting and tracking capabilities in the field of Neuromorphic Vision.

The developed framework can be used to test and evaluate several Neuromorphic Vision algorithms using conventional datasets in Computer Vision. We recognize the limitation of our system in approaches that require the extreme temporal resolution of real neuromorphic devices. However, many applications can benefit from the process of converting traditional information into neuromorphic information.

Conflict of Interest

The authors declare that they have no conflict of interest.

Acknowledgements

We thank the funding agencies CNPq, CAPES and FAPEMIG for the support.

References

1. Diamond Alan, Nowotny Thomas, Schmuker Michael. Comparing neuromorphic solutions in action: implementing a bio-inspired solution to a benchmark classification task on three parallel-computing platforms *Frontiers in neuroscience.* 2016;9:491.
2. Osborn Luke E, Dragomir Andrei, Betthauser Joseph L, et al. Prosthesis with neuromorphic multilayered e-dermis perceives touch and pain *Science robotics.* 2018;3:eaat3818.
3. Buccelli Stefano, Bornat Yannick, Colombi Ilaria, et al. A neuromorphic prosthesis to restore communication in neuronal networks *iScience.* 2019;19:402–414.
4. Delbrück Tobi, Linares-Barranco Bernabe, Culurciello Eugenio, Posch Christoph. Activity-driven, event-based vision sensors in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*:2426–2429IEEE 2010.
5. Brandli Christian, Berner Raphael, Yang Minhao, Liu Shih-Chii, Delbruck Tobi. A 240× 180 130 db 3 μs latency global shutter spatiotemporal vision sensor *IEEE Journal of Solid-State Circuits.* 2014;49:2333–2341.
6. Paredes-Vallés Federico, Scheper Kirk Yannick Willehm, De Croon Guido Cornelis Henricus Eugene. Unsupervised learning of a hierarchical spiking neural network for optical flow estimation: From events to global motion perception *IEEE transactions on pattern analysis and machine intelligence.* 2019.
7. Delbruck Tobi. Frame-free dynamic digital vision in *Proceedings of Intl. Symp. on Secure-Life Electronics, Advanced Electronics for Quality Life and Society*:21–26Citeseer 2008.
8. Delbruck Tobi, Lichtsteiner Patrick. Fast sensory motor control based on event-based hybrid neuromorphic-procedural system in *2007 IEEE international symposium on circuits and systems*:845–848IEEE 2007.
9. Gallego Guillermo, Lund Jon EA, Mueggler Elias, Rebecq Henri, Delbruck Tobi, Scaramuzza Davide. Event-based, 6-DOF camera tracking from photometric depth maps *IEEE transactions on pattern analysis and machine intelligence.* 2017;40:2402–2412.
10. Ni Zhenjiang, Ieng Sio-Hoi, Posch Christoph, Régnier Stéphane, Benosman Ryad. Visual tracking using neuromorphic asynchronous event-based cameras *Neural computation.* 2015;27:925–953.
11. Tan Cheston, Lallee Stephane, Orchard Garrick. Benchmarking neuromorphic vision: lessons learnt from computer vision *Frontiers in neuroscience.* 2015;9:374.
12. Mueggler Elias, Rebecq Henri, Gallego Guillermo, Delbruck Tobi, Scaramuzza Davide. The Event-Camera Dataset: Event-based Data for Pose Estimation, Visual Odometry, and SLAM
13. Dubuisson Séverine, Gonzales Christophe. A survey of datasets for visual tracking *Machine Vision and Applications.* 2016;27:23–52.
14. About the VOT 2014 dataset at shorturl.at/bdvNZ
15. Visual Tracker Benchmark at shorturl.at/rtAJ1
16. Sunderrajan Santhoshkumar, Jagadeesh Vignesh, Manjunath BS. Robust Multiple Camera Tracking with Spatial And Appearance Contexts 2015.
17. Santner Jakob, Leistner Christian, Saffari Amir, Pock Thomas, Bischof Horst. PROST: Parallel robust online simple tracking in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*:723–730IEEE 2010.
18. 2D MOT 2015 at shorturl.at/bhwFN
19. Serra Jean. *Image analysis and mathematical morphology*. Academic Press, Inc. 1983.
20. Digabel H, Lantuéjoul Christian. Iterative algorithms in *Proc. 2nd European Symp. Quantitative Analysis of Microstructures in Material Science, Biology and Medicine*;19:8Stuttgart, West Germany: Riederer Verlag 1978.
21. Vincent Luc, Soille Pierre. Watersheds in digital spaces: an efficient algorithm based on immersion simulations *IEEE Transactions on Pattern Analysis & Machine Intelligence.* 1991:583–598.
22. Preim Bernhard, Botha Charl. Chapter 4—Image Analysis for Medical Visualization *Visual Computing for Medicine, 2nd ed.; Preim, B., Botha, C., Eds.* 2014:111–175.
23. Mangan Alan P, Whitaker Ross T. Partitioning 3D surface meshes using watershed segmentation *IEEE Transactions on Visualization and*

Computer Graphics. 1999;5:308–321.

24. Nguyen Hieu Tat, Worring Marcel, Van Den Boomgaard Rein. Watersnakes: Energy-driven watershed segmentation *IEEE Transactions on Pattern Analysis and Machine Intelligence.* 2003;25:330–342.
25. Li Jiangbo, Luo Wei, Wang Zheli, Fan Shuxiang. Early detection of decay on apples using hyperspectral reflectance imaging combining both principal component analysis and improved watershed segmentation method *Postharvest biology and technology.* 2019;149:235–246.
26. Rabbani Arash, Ayatollahi Shahab, Kharrat Riyaz, Dashti Nader. Estimation of 3-D pore network coordination number of rocks from watershed segmentation of a single 2-D image *Advances in Water Resources.* 2016;94:264–277.
27. Rogowska Jadwiga. Overview and fundamentals of medical image segmentation *Handbook of medical imaging, processing and analysis.* 2000:69–85.
28. Hahn Horst Karl, Peitgen Heinz-Otto. IWT-Interactive Watershed Transform: A hierarchical method for efficient interactive and automated segmentation of multidimensional grayscale images in *Medical Imaging 2003: Image Processing*;5032:643–653International Society for Optics and Photonics 2003.
29. Kuhnigk Jan-Martin, Hahn Horst, Hindennach Milo, Dicken Volker, Krass Stefan, Peitgen Heinz-Otto. Lung lobe segmentation by anatomy-guided 3D watershed transform in *Medical imaging 2003: image processing*;5032:1482–1490International Society for Optics and Photonics 2003.
30. Orchard Garrick, Jayawant Ajinkya, Cohen Gregory K, Thakor Nitish. Converting static image datasets to spiking neuromorphic datasets using saccades *Frontiers in neuroscience.* 2015;9:437.
31. Zhu Alex Zihao, Thakur Dinesh, Özaslan Tolga, Pfrommer Bernd, Kumar Vijay, Daniilidis Kostas. The multivehicle stereo event camera dataset: An event camera dataset for 3D perception *IEEE Robotics and Automation Letters.* 2018;3:2032–2039.
32. Mitrokhin Anton, Ye Chengxi, Fermuller Cornelia, Aloimonos Yiannis, Delbruck Tobi. EV-IMO: Motion segmentation dataset and learning pipeline for event cameras *arXiv preprint arXiv:1903.07520.* 2019.
33. Mitrokhin Anton, Fermüller Cornelia, Parameshwara Chethan, Aloimonos Yiannis. Event-based moving object detection and tracking in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*:1–9IEEE 2018.
34. Pérez-Carrasco José Antonio, Zhao Bo, Serrano Carmen, et al. Mapping from frame-driven to frame-free event-driven vision systems by low-rate rate coding and coincidence processing–application to feedforward ConvNets *IEEE transactions on pattern analysis and machine intelligence.* 2013;35:2706–2719.
35. N-CARS at prophesee.ai/2018/03/13/dataset-n-cars/
36. Hu Yuhuang, Liu Hongjie, Pfeiffer Michael, Delbruck Tobi. DVS benchmark datasets for object tracking, action recognition, and object recognition *Frontiers in neuroscience.* 2016;10:405.
37. Katz Matthew L, Nikolic Konstantin, Delbruck T. Live demonstration: Behavioural emulation of event-based vision sensors in *2012 IEEE International Symposium on Circuits and Systems*:736–740IEEE 2012.
38. Bi Yin, Andreopoulos Yiannis. PIX2NVS: Parameterized conversion of pixel-domain video frames to neuromorphic vision streams in *2017 IEEE International Conference on Image Processing (ICIP)*:1990–1994IEEE 2017.

Author: Eduardo Borges Gouveia
Institute: Biomedical Engineering Laboratory - BIOLAB
Street: Av. João Naves de Ávila, 2121
City: Uberlândia
Country: Brazil
Email: eduardoborgesgouveia@ufu.br