POST-STIMULUS TIME-DEPENDENT EVENT DESCRIPTOR

S. Harrigan, S. Coleman, D. Kerr, P. Yogarajah

Faculty of Computing,
Engineering and Built Environment
Ulster University
Northern Ireland, UK

Z. Fang, C. Wu

Faculty of Robot Science and Engineering
Northeastern University
Shenyang, China

ABSTRACT

Event-based image processing is a relatively new domain in the field of computer vision. Much research has been carried out on adapting event-based data to comply with established techniques from frame-based computer vision. On the contrary, this paper presents a descriptor which is designed specifically for direct use with event-based data and therefore can be considered to be a pure event-based vision descriptor as it only uses events emitted from event-based vision devices without transforming the data to accommodate frame-based vision techniques. This novel descriptor is known as the Post-stimulus Time-dependent Event Descriptor (P-TED). P-TED is comprised of two features extracted from event data which describe motion and the underlying pattern of transmission respectively. Furthermore a framework is presented which leverages the P-TED descriptor to classify motions within event data. This framework is compared against another state-of-the-art event-based vision descriptor as well as an established frame-based approach.

Index Terms— Bio-inspired, Neuromorphic, Motion Recognition, Multi-dimensional Signal Processing, Computer Vision

1. INTRODUCTION

Event-based vision sensors are considerably different from the frame-based vision sensors widely deployed today. Frame-based vision sensors transmit data synchronously in the form of 2D frames representing the photodetector pixel array signals at the time of capture. Other than the biological-inspiration of event-based vision devices, the main divergence between frame-based and event-based devices is the synchronous and asynchronous nature of data transmission. The synchronized transmission in frame-based devices allows for the capture of static scene details but contributes to latency. Event-based devices overcome this transmitting data corresponding to the location where a luminance change is detected, asynchronously at event resolution times (e.g. 15 microseconds) [1].

Research into processing event-based vision data initially focussed on feature extraction such as the detection of edges and corners [2], and often creates a standard frame-based image from event data to achieve this. Another common processing approach used throughout event-based vision processing is the use of the time-surfaces [3]. The asynchronous nature of event data means that most event data processing techniques will need to rely on some form of memory infrastructure in order to interpret the data beyond individual events which the time-surface fulfils. The time-surface is designed to retain prominent spatio-temporal information using a multidimensional lattice. Each cell of the lattice maps to a event pixel within a device and retains the temporal information of the most recent event activity from the pixel. The time-surface is simply a means of converting event data into a frame (known as an event-frame) at event resolutions and many variations can be produced due to the innate extension property (e.g. speed in-variance [4]) of the lattice.

Other approaches to event-data processing commonly rely on machine learning [5], clustering [6] and/or converting the data to frames or contrast maps in combinations with established image processing techniques such as the Harris corner detector [7] or FAST [8]. With the growth of machine learning in the field of computer vision, event-based image processing currently mainly involves the use of some form of machine learning technique such as a Support Vector Machine (SVM) [9], Random Forest [4], a Convolutional Neural Network [10], a Spiking Neural Network (SNN) [11] [12], Recurrent Neural Network (RNN) [13] or Deep Neural Networks (DNN) [14]. In [9] a novel descriptor based on retinal transform theory has been used as data for a machine learning pipeline in a classification problem setting.

Building on [15] we present a novel descriptor designed specifically for event-based image processing known as Post-stimulus Time-dependent Event Descriptor (P-TED). We demonstrate that this can be combined with a matching framework for robot motion. The presented framework is evaluated with a state-of-the-art approach across two established datasets. The P-TED descriptor and the presented framework are found to be an efficient but robust means of

representing motions in event data. The remainder of the paper is as follows: Section 2 describes the novel P-TED approach in detail and Section 3 presents the classification framework developed. Section 4 presents the various experiments used and the performance evaluation. Finally the work is concluded in Section 5.

2. POST-STIMULUS TIME-DEPENDENT EVENT DESCRIPTOR (P-TED)

Event-based vision sensors, such as the Dynamic Vision Sensor (DVS) [16], are inspired by the neural processing systems in the retina. When a change in luminance is detected, an event is triggered at that particular time t along with the specific sensor in the pixel array l where the luminance change is detected. A polarity mechanism p which indicates if the change in luminance was positive (luminance increased indicating a brightening of the region) or negative (indicating a darkening of the region) is utilised. An event e can then be expressed as $e \in \langle t, l, p \rangle$ where $l = \langle x, y \rangle$ with x, y corresponding to a pixel location within a 2-D pixel array. A stream of events at any given time period can be denoted as $S(e_1, ..., e_m)$.

The P-TED descriptor operates over a subset R of event data S where the subset contains pairs of correlated events r_i from the time-surface T. The pairs of correlated events r_i are considered to be events which have spatial-temporal closeness to other events within S. To correlate events, a Moore neighbourhood [17] is used on a time-surface (a memory infrastructure which is used to track events in the past using a 2-D lattice) [3]. A Moore neighbourhood is a 2-D 3×3 lattice with a centre cell C and eight surrounding spatial neighbours which are referred to by their cardinal and intercardinal position relative to C (N, NE, E, SE, S, SW, W, NW). Figure 1 illustrates a Moore neighbourhood with it's cardinal and intercardinal neighbours. Let the Moore neighbourhood be H. Then, if a neighbour e_n, where $n = \{H_N, H_{NE}, \ldots, H_{NW}\}$ and $n \neq H_C$, of the centre event e_c contains an event, then that event is deemed to be spatially close to e_c and is therefore added to the time-surface T. To determine the temporal closeness δ_t of events, e_t, in T the timestamp of the centre event e_c is compared with each event e_t such that $\delta_t = e_c - e_t$. The event e_t, which when compared with e_c, is closest in time δ_t and less than a threshold Δ (e.g. 5 microseconds), it is deemed to satisfy the temporal constraint and be temporally closest. If an event e_t is determined to satisfy the spatial-temporal constraints, it is paired with e_c to form r_i and added to the set of correlated events R. If no e_t can be found to satisfy the spatial-temporal closeness constraints then e_c remains on the time-surface T.

The novel Post-stimulus Time-dependent Event Descriptor (P-TED) is a pure event-based vision descriptor which is applied directly to the set R of correlated events. P-TED consists of two features, motion direction V and pattern G (the pattern of event pixels excitation over time). Motion direction V is discussed in detailed in Section 2.1 and pattern G is discussed in detail in Section 2.2.

2.1. Motion Direction Feature

The motion direction vector V represents the number of correlated pairs r_i in each of the eight cardinal and intercardinal directions of the Moore neighbourhood. For each pair of correlated events $r_i = \{e_c, e_t\}$ we calculate the angle between e_c and e_t and determine which of the eight cardinal and intercardinal direction bins it corresponds to. Therefore the values in V correspond to the number of correlated pairs r_i in each of the directions (N, NE, E, SE, S, SW, W, NW).

For example, if it is found that there are 12 N-directional correlated pairs, 4 NE-directional correlated pairs and 7 E-directional correlated pairs, then $V = \{12, 4, 7, 0, 0, 0, 0, 0\}$ (observations occur in a clockwise fashion starting at N). The use of the Moore neighbourhood allows us to express V as a motion direction which in this example the event motion is predominantly to the north with a slightly eastern bearing in this example.

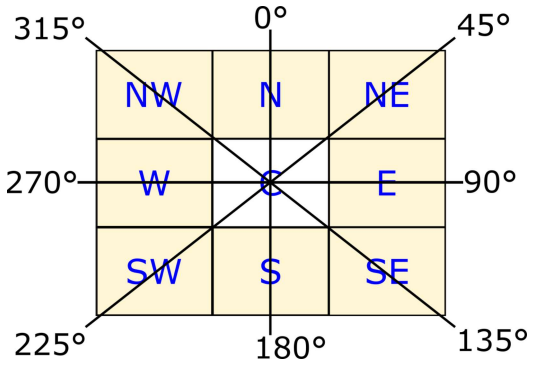

Fig. 1. Moore neighbhourhood

2.2. Pattern Feature

The pattern vector G is a rate encoding technique [18] which determines the number of event pairs r_i that occur within a specific time bin. For any given stream of events S, the appropriate bin size β for the pattern vector G needs to be determined. In order to do so we use the approach from [19] in which G is initially divided into γ bins of size β and the number N_j of event pairs r_i that are in each bin j. We then calculate the mean f and variance d of N_j within G, thus we compute the cost function

$$(2f - d)/\beta^2 \tag{1}$$

Through an iterative process of varying the bin size β we minimise the cost function (1) thus determining the optimal β. Hence, computing the number of events that occur within

each γ results in a pattern feature G of the event activity. This is considered analogous to the established neuroscience finding of time-to-first spike encoding [20].

3. CLASSIFICATION FRAMEWORK

A novel classification framework is designed that utilises P-TED. We compute a set of corresponding P-TEDs for the Motions of Interest (MoI) M which we wish to classify. The framework makes use of a sliding window, J, which has the same length as the MoI, operating over a live stream of event data LS, and computes a P-TED for each window. This framework is illustrated in Figure 2.

For simplicity, the MoI P-TED feature vectors are denoted as $M(V)$ and $M(G)$ for motion direction and pattern respectively. Similarly, within a given window J the feature vectors are denoted as $J(V)$ and $J(G)$. To determine the similarity between the MoI and J we consider the motion direction vector pair and the pattern vector pair respectively. The similarity between the motion direction vectors is computed as:

$$Sim(V) = \frac{\sum_{i=1}^{8} \begin{cases} 0 & M(V) \neq J(V) \\ 1 & M(V) = J(V) \end{cases}}{8} \tag{2}$$

resulting in a continuous value, between 0 and 1 where 1 is a perfect match, representing the similarity between the pattern vector of MoI and the pattern vector of the current window. The similarity between the pattern vectors is computed as:

$$Sim(G) = \mid \sum_{i=0}^{\gamma} \frac{J(G_i)}{M(G_i) + \epsilon} \mid \tag{3}$$

where ϵ is an appropriately small number to avoid division by zero. The outputs from equations (2) and (3) are used to compute the overall similarity U such that:

$$U = \frac{Sim(G) \cdot Sim(V)}{min(Sim(G), Sim(V)) + \epsilon} \tag{4}$$

where ϵ is an appropriately small number to avoid division by zero. The similarity measure U is an overall numerical representation of how similar the MoI is to J within the event data stream. Therefore, this framework consists of two key components, P-TED (Section 2) and the overall similarity measure.

4. PERFORMANCE EVALUATION

In order to determine the accuracy and efficiency of the proposed P-TED framework, we compare with a similar state-of-the-art event data based framework known as Distribution Aware Retinal Transform (DART) [9].

The DART descriptor encodes the event data using a log-polar grid to simulate the distribution of photo-receptor cones in the primate fovea [21]. The DART framework utilises a Support Vector Machine (SVM) for motion classification. Therefore the DART framework is composed of two key components, the DART descriptor and the SVM.

The performance of P-TED and DART is compared using the MNIST-DVS [22] and the CIFAR10-DVS [23] datasets. MNIST-DVS is a neuromorphic version of the popular MNIST dataset which contains 70,000 handwriting samples of 10 classes representing a digit range of 0 - 9. The MNIST-DVS contains 30,000 event vision sensor responses to 30,000 handwriting samples. The controlled motion of the MNIST handwriting sample is presented on a screen to enable responses to be captured by a 128 x 128 DVS device. Similarly, the CIFAR10-DVS is an event-based version of the CIFAR10. It consists of 10,000 event data streams where 1,000 streams are used for each of the 10 classes (airplanes, cars, birds, cats, deer, dogs, frogs, horses, ships and trucks). Figure 3 shows example images and event frames for the truck, automobile and frog classes for both CIFAR10 and CIFAR10-DVS respectively.

Initially we directly compare the P-TED descriptor with the DART descriptor within the DART framework (SVM). Using the MNIST-DVS and the CIFAR10-DVS we calculate both the recognition accuracy and the average run-time. Results are presented in Table 1. In Table 1 we can see that the P-TED descriptor provides the highest accuracy performance using both datasets. The average run-time the DART framework is 10.3 ms.

Table 1. Recognition accuracy results from using the DART framework

DATASET	P-TED(%)	DART(%)	Time (ms)
MNIST-DVS	98.31	97.95	10.3
CIFAR10-DVS	66.90	65.78	

For completeness we then compare the P-TED descriptor with the DART descriptor within the P-TED framework (Similarity measure U, Section 3). The results are presented in Table 2 where we can see that again the P-TED descriptor provides the highest performance accuracy for both datasets, however the overall performance accuracy using the P-TED framework is lower than that presented in Table 1. The average runtime for the P-TED framework is 2.7 ms demonstrating increased efficiency when calculating a similarity measure compared with using a clasification technique such as the SVM.

Table 2. Recognition accuracy results from using the P-TED framework

DATASET	P-TED(%)	DART(%)	Time (ms)
MNIST-DVS	91.73	86.74	2.7
CIFAR10-DVS	58.55	51.74	

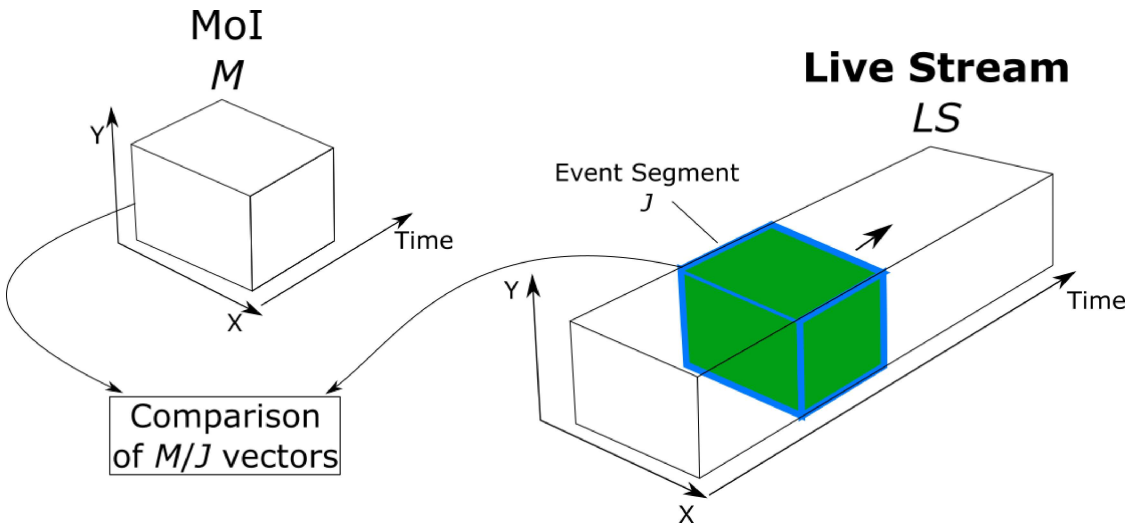

Fig. 2. An illustration of the P-TED framework showing the MoI M, the live stream of events LS with an event segment J extracted using a sliding window over space-time and the comparison of the M and J vectors.

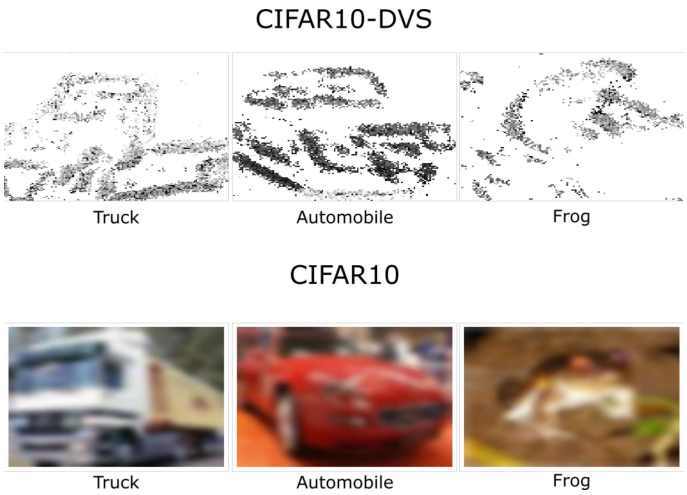

Fig. 3. Examples from CIFAR10 and CIFAR10-DVS

Whilst the DART framework produced the highest results using both P-TED and DART descriptors it must be noted that using the P-TED descriptor in either framework produced the highest accuracy score. The P-TED framework provides a strong level of accuracy with a $4\times$ speed-up compared with the DART framework. An explanation for this contrasting difference is that DART is heavily reliant on the spatial relationship of events within the correlation and transform stages of the framework, P-TED only places the spatial constraint on correlation and strives to represent the motion direction and pattern components within the event data separately instead of binding them as DART does.

5. CONCLUSION

This paper presents a novel approach to event-based data processing known as Post-stimulus Time-dependent Event Descriptor (P-TED), comprising of two feature vectors representing motion direction and pattern. The P-TED descriptor is combined with a novel similarity measure which enables us to classify a range of different motions. This is demonstrated by the performance evaluation which was conducted using two well known event-based datasets. The P-TED descriptor demonstrated superior performance compared with the DART descriptor [9]. Analysis has shown us the pattern histogram is affected by scale, thus future work will be focussed on making the P-TED framework scale invariant.

6. REFERENCES

[1] Elias Mueggler, Basil Huber, and Davide 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.

[2] Xavier Clady, Sio-Hoi Ieng, and Ryad Benosman, "Asynchronous event-based corner detection and matching," *Neural Networks*, vol. 66, pp. 91–106, 2015.

[3] Xavier Lagorce, Garrick Orchard, Francesco Galluppi, Bertram E. Shi, and Ryad B. Benosman, "HOTS: A Hierarchy of Event-Based Time-Surfaces for Pattern Recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 39, no. 7, pp. 1346–1359, 2017.

[4] Jacques Manderscheid, Amos Sironi, Nicolas Bourdis, Davide Migliore, and Vincent Lepetit, "Speed invariant time surface for learning to detect corner points with event-based cameras," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019, pp. 10245–10254.

[5] Iulia-Alexandra Lungu, Federico Corradi, and Tobi Delbrück, "Live demonstration: Convolutional neural network driven by dynamic vision sensor playing roshambo," in *2017 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2017, pp. 1–1.

[6] Ewa Piatkowska, Ahmed Nabil Belbachir, Stephan Schraml, and Margrit Gelautz, "Spatiotemporal multiple persons tracking using dynamic vision sensor," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2012, pp. 35–40.

[7] Valentina Vasco, Arren Glover, and Chiara Bartolozzi, "Fast event-based Harris corner detection exploiting the advantages of event-driven cameras," in *IEEE International Conference on Intelligent Robots and Systems*. oct 2016, vol. 2016-Novem, pp. 4144–4149, IEEE.

[8] Elias Mueggler, Chiara Bartolozzi, and Davide Scaramuzza, "Fast Event-based Corner Detection," in *British Machine Vis. Conf. (BMVC)*, 2017, vol. 1, pp. 1–11.

[9] Bharath Ramesh, Hong Yang, Garrick Orchard, Ngoc Anh Le Thi, and Cheng Xiang, "Dart: distribution aware retinal transform for event-based cameras," *arXiv preprint arXiv:1710.10800*, 2017.

[10] Hongjie Liu, Diederik Paul Moeys, Gautham Das, Daniel Neil, Shih-Chii Liu, and Tobi Delbrück, "Combined frame-and event-based detection and tracking," in *2016 IEEE International Symposium on Circuits and Systems (ISCAS)*. IEEE, 2016, pp. 2511–2514.

[11] Luis Alejandro Camuñas-Mesa, Teresa Serrano-Gotarredona, and Bernabé Linares-Barranco, "Event-driven sensing and processing for high-speed robotic vision," in *2014 IEEE Biomedical Circuits and Systems Conference (BioCAS) Proceedings*. IEEE, 2014, pp. 516–519.

[12] Sambit Mohapatra, Heinrich Gotzig, Senthil Yogamani, Stefan Milz, and Raoul Zollner, "Exploring deep spiking neural networks for automated driving applications," *arXiv preprint arXiv:1903.02080*, 2019.

[13] Henri Rebecq, René Ranftl, Vladlen Koltun, and Davide Scaramuzza, "High speed and high dynamic range video with an event camera," *arXiv preprint arXiv:1906.07165*, 2019.

[14] Ana I Maqueda, Antonio Loquercio, Guillermo Gallego, Narciso García, and Davide Scaramuzza, "Event-based vision meets deep learning on steering prediction for self-driving cars," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 5419–5427.

[15] Shane Harrigan, Dermot Kerr, Sonya Coleman, Pratheepan Yogarajah, Zheng Fang, and Chengdong Wu, "Neuromorphic event-based space-time template action recognition," 8 2018, Irish Machine Vision and Image Processing Conference, IMVIP ; Conference date: 29-08-2018 Through 31-08-2018.

[16] Patrick Lichtsteiner, Christoph Posch, and Tobi Delbruck, "A 128 x 128 120 dB 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[17] Edward F Moore, "Machine models of self-reproduction," in *Proceedings of symposia in applied mathematics*. American Mathematical Society New York, 1962, vol. 14, pp. 17–33.

[18] Markus Weber, "Single unit recordings–peristimulus time histograms (psths)," in *Handbook of Clinical Neurophysiology*, vol. 4, pp. 349–358. Elsevier, 2004.

[19] Hideaki Shimazaki and Shigeru Shinomoto, "A method for selecting the bin size of a time histogram," *Neural computation*, vol. 19, no. 6, pp. 1503–1527, 2007.

[20] T. Gollisch and M. Meister, "Rapid Neural Coding in the Retina with Relative Spike Latencies," *Science*, vol. 319, no. 5866, pp. 1108–1111, 2008.

[21] Eric L Schwartz, "Spatial mapping in the primate sensory projection: analytic structure and relevance to perception," *Biological cybernetics*, vol. 25, no. 4, pp. 181–194, 1977.

[22] Teresa Serrano-Gotarredona and Bernabé Linares-Barranco, "Poker-dvs and mnist-dvs. their history, how they were made, and other details," *Frontiers in neuroscience*, vol. 9, pp. 481, 2015.

[23] Hongmin Li, Hanchao Liu, Xiangyang Ji, Guoqi Li, and Luping Shi, "Cifar10-dvs: an event-stream dataset for object classification," *Frontiers in neuroscience*, vol. 11, pp. 309, 2017.