# Neuromorphic Event-based Action Recognition

S. Harrigan [1], S. Coleman [1], D. Kerr [1], P. Yogarajah [1], Z. Fang [2], C. Wu [2]

[1] *Intelligent Systems Research Centre, Ulster University, Northern Ireland, United Kingdom*
[2] *College of Information Science and Engineering, Northeastern University, Shenyang, China*
*{harrigan-s, d.kerr, sa.coleman, p.yogarajah}@ulster.ac.uk {fangzheng, wuchengdong}@mail.neu.edu.cn*

**Abstract**

An action can be viewed as spike trains or streams of events when observed and captured by neuromorphic imaging hardware such as the iniLabs DVS128. These streams are unique to each action enabling them to be used to form descriptors. This paper describes an approach for detecting specific actions based on space-time template matching by forming such descriptors and using them as comparative tools. The developed approach is used to detect symbols from the popular RoShambo (rock, paper and scissors) game. The results demonstrate that the developed approach can be used to correctly detect the motions involved in producing RoShambo symbols.

**Keywords:** action detection, neuromorphic processing, space-time, game, RoShambo.

## 1 Introduction

Event sensors, such as the Dynamic Vision Sensor (DVS) [Lichtsteiner et al., 2008] are shifting the way in which the capture of visual data occurs. Event sensors work through the dynamics of scene activity instead of a fixed capture rate. For example, the DVS is an alternative to traditional frame-based cameras and works by emitting 'events' of luminance change asynchronously rather than the synchronous frame-based approach were pixels are used to form frames. Event camera sensors such as the DVS work independently of the other sensors in their familial array, working asynchronously. When a DVS sensor reports a change in luminance it emits three pieces of information; spatial occurrence coordinates $(x,y)$ relating to the sensor which emitted the luminance change report, a polarity value $p$ relating to whether the luminance change is an increase or decrease in value where $p \in \{+1,-1\}$ and a timestamp $t$ relating to the time emitted by the sensor; in summary each event $e \in \langle t, \langle x, y\rangle, p\rangle$. Event cameras often consume less energy, produce sparse data and have low transmission latency compared with their frame generating counterparts [Brandli et al., 2014][Lichtsteiner et al., 2008].

The detection of motions is often a desired attribute of systems which involve monitoring. Techniques for the detection of motions are both well understood and varied in terms of frame-based vision processing. They range from observing the effect intensity change has on the underlying motion fields [Shechtman and Irani, 2005] to converting the motions into shapes and measuring different properties of those shapes [Blank et al., 2005]. Event-based vision processing, being a recent newly developed field, does not possess such well understood methods. Recent works involving event-based vision have included using both frame and event data together to track corners [Tedaldi et al., 2016], classifying objects using histograms of averaged time surfaces [Sironi et al., 2018] and focusing on the use of machine learning techniques [Maqueda et al., 2018, Thiele et al., 2018].

## 2 Event-based Motion Matching

The proposed approach (Figure 1) uses an iniLabs DVS128 [Lichtsteiner et al., 2008] which signals luminance change at the pixel level of an observed scene. There are three main steps towards identifying an action in this approach. Firstly, the Euclidean distance between each event in the sequence is calculated, those which exceed a distance value (usually around 2 pixels in space) are not considered relevant to the action being detected. In order to estimate if two events are correlated we use a 3x3 spatial-temporal matrix centered on the pixel location of the current event which keeps a map of the last event time at a particular pixel location. Events that are not correlated are isolated thus not added to the calculation but are still recorded in the map for future calculations. Both the template $T$ (a collection of events of a controlled recording representing the desired action) and the stream of unknown events $V$ contain correlated events which have been processed in this manner. Secondly,

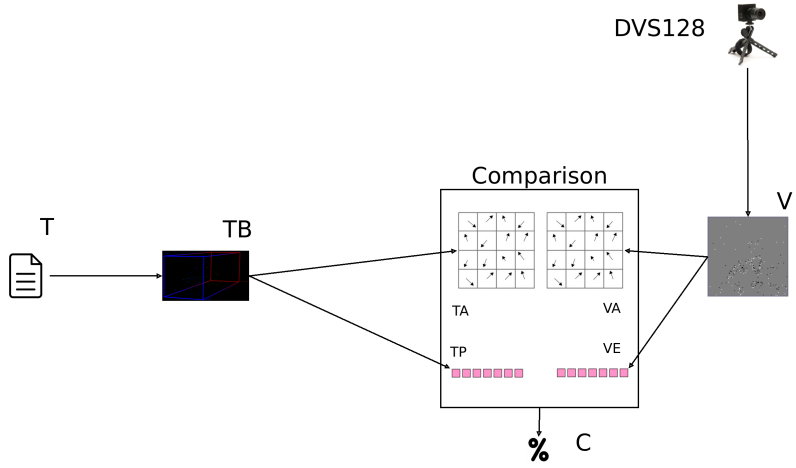


Figure 1: Stages of proposed approach

the orientation of each event in sequence within a template $T$ containing a list of events $TB$ is calculated and an angle orientation bin $T_A$ position, representing the calculated angle, is increased incrementally for each occurrence of that angle within $T$. Additionally, the pattern inside $TB$ is extracted linearly as $T_P$. Both $T_A$ and $T_P$ can be considered to be descriptors of $T$. The same approach is taken for each window of $V$ resulting in angle orientation bins $V_A$. The key assumption for both $T_A$ and $V_A$ is that the only events being processed after the first step are events which are associated with each other.

A segment $V_S$ of $V$ is compared with $T$ using a temporal sliding window, where the length of the sliding window is equal to the temporal length of $T$ (usually around 1000 microseconds but this is dependent on the motion being captured). The proposed system calculates an angle orientation bin for $V_S$, $T_P$ is compared with the pattern of events $V_E$ of $V$ as they are presented. At the end of the window, the similarity between $T_A$ and $V_A$ is calculated by comparing the difference between the sum of the bin of $V_A$ and $T_A$; a high similarity value denoted as $R_A$ indicates if $T_A$ and $V_A$ are similar actions. Comparing $V_E$ with $T_P$ produces a similarity score RP that determines close $V_E$ is to $T_P$, a high value of $R_P$ indicates that the pattern of events in $V_E$ are very similar to $T_P$ and thus gives an indication of how similar the two induced motion fields are to one another. We calculate $R_A$ and $R_P$ respectively as:

$$R_G = \| \sum_{i=0}^{\gamma} \frac{V_{P_i}}{T_{P_i} + \epsilon} \| \tag{1}$$

$$R_P = \frac{\sum_{i=0}^{n} \begin{cases} 0 & V_{P_i} \neq T_{P_i} \\ 1 & V_{P_i} = T_{P_i} \end{cases}}{n} \tag{2}$$

where $n$ is the number of events within the template, $W_i$ is the value of the bin in the window angle orientation descriptor at $i$, $T_i$ is the value of the bin at $i$ in the template angle orientation descriptor, $V_{P_i}$ is the polarity

value in the pattern descriptor of the window at $i$, $T_{P_i}$ is the polarity value at $i$ in the template pattern descriptor and $\epsilon$ is set appropriately small to avoid division by zero. Thirdly, the similarity score $C$ which compares an observed action is to the action template is calculated using Equation 3.

$$C = \frac{R_P \cdot R_A}{min(R_P, R_A) + \epsilon} \tag{3}$$

where $R_A$ and $R_P$ are obtained from Equation 1 and Equation 2 respectively and $\epsilon$ avoids division by zero. A key assumption of the approach is that packets of events within the window belong to a monotonically increasing sequence in terms of each event's occurrence in the observed scene. This is reinforced through the use of the 3x3 spatial-temporal neighbourhood.

Figure 2 shows the RoShambo motions and their relationship to each other, for example, rock is beaten by paper. The designed system made use of templates of the three distinctive motions of forming rock, paper and scissors, with these being the template of each individual symbol throughout this experiment.

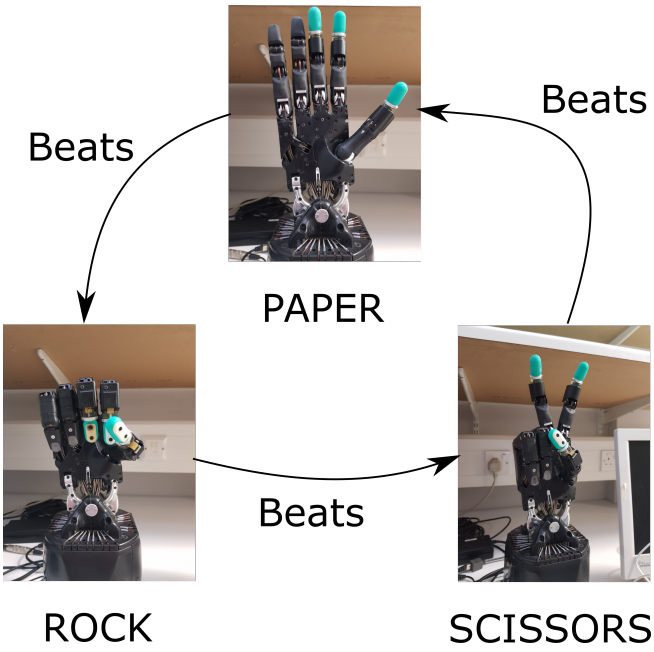


Figure 2: An illustration of RoShambo symbols and the relationship between each.

## 3 Experimental Results

The proposed approach was applied to the detection of RoShambo symbols using a Shadow Hand robot [Kochan, 2005], the experiment was inspired by demonstrative work. The detection output of the approach is determined to be the template producing the highest similarity score (Equation 3) with 10 milliseconds of the symbol being formed. The system then outputs the appropriate response to the RoShambo symbol (for example, if the Shadow Hand produce scissors, the system would output rock). If the approach output the incorrect response to the Shadow Hand RoShambo symbol it was counted as a lose. The Euclidean distance between events was limited to 10 pixels and $\epsilon$ was set to 0.0000001 (which was chosen to avoid division by zero).

The results of two experimental runs are shown in Table 1 and demonstrate that the proposed method can detect the RoShambo forming motions via the use of neuromorphic cameras and the proposed template matching approach. The results indicate that the developed approach's pattern and motion descriptors are capable of determining, to a high accuracy, very fine action motions.

| RoShambo Symbol | Occurrence | Detected | Error |
|---|---|---|---|
| *Paper* | 5 | 5 | 0 |
| *Rock* | 5 | 4 | 1 |
| *Scissors* | 5 | 5 | 0 |

Table 1: RoShambo experiment results showing the RoShambo symbol occurrence, the number of times the occurrence was detected and the number of times the occurrence was missed.

# 4 Conclusion

The results of the RoShambo experiment show that this system can accurately detect the motions used to form RoShambo symbols. An area for improvement for future research is the application of a multi-scale framework which is designed for event-based data.

The ability to detect an action using a generic template means that the requirements for configuring and maintaining such a system based on a motion are reduced in the long term. Although a generic template of an action is not always capable of capturing every action, the preliminary experiments provide sufficient evidence that this type of fine grain action can be readily detected using an event-based approach.

# References


[Blank et al., 2005] Blank, M., Gorelick, L., Shechtman, E., Irani, M., and Basri, R. (2005). Actions as Space-Time Shapes. In *IEEE International Conference on Computer Vision*, volume 2, pages 1395–1402. IEEE.

[Brandli et al., 2014] Brandli, C., Berner, R., Yang, M., Liu, S. C., and Delbruck, T. (2014). A 240 x 180 130 dB 3 $\mu$s latency global shutter spatiotemporal vision sensor. *IEEE Journal of Solid-State Circuits*, 49(10):2333–2341.

[Kochan, 2005] Kochan, A. (2005). Shadow delivers first hand. *Industrial robot: an international journal*, 32(1):15–16.

[Lichtsteiner et al., 2008] Lichtsteiner, P., Posch, C., and Delbruck, T. (2008). A 128 x 128 120 dB 15 $\mu$s latency asynchronous temporal contrast vision sensor. *IEEE Journal of Solid-State Circuits*, 43(2):566–576.

[Maqueda et al., 2018] Maqueda, A. I., Loquercio, A., Gallego, G., García, N., and Scaramuzza, D. (2018). 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*, pages 5419–5427.

[Shechtman and Irani, 2005] Shechtman, E. and Irani, M. (2005). Space-time behavior based correlation. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1:405–412.

[Sironi et al., 2018] Sironi, A., Brambilla, M., Bourdis, N., Lagorce, X., and Benosman, R. (2018). HATS: Histograms of Averaged Time Surfaces for Robust Event-based Object Classification. *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1731–1740.

[Tedaldi et al., 2016] Tedaldi, D., Gallego, G., Mueggler, E., and Scaramuzza, D. (2016). Feature detection and tracking with the dynamic and active-pixel vision sensor (DAVIS). In *2016 2nd International Conference on Event-Based Control, Communication, and Signal Processing, EBCCSP 2016 - Proceedings*, pages 1–7. IEEE.

[Thiele et al., 2018] Thiele, J. C., Bichler, O., and Dupret, A. (2018). A timescale invariant stdp-based spiking deep network for unsupervised online feature extraction from event-based sensor data. In *2018 International Joint Conference on Neural Networks (IJCNN)*, pages 1–8. IEEE.