MEng Individual Project

Imperial College London

Department of Electrical and Electronic Engineering

Event Analysis for Classification of Neuropmorhic Data

Author:
Tejas Dandawate

Supervisor:
Prof. Pier Luigi Dragotti

Second Marker:
Prof. Patrick A. Naylor

May 28, 2022

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivations

Neuromorphic data is obtained from event-based cameras that differ from traditional frame-based cameras in that they asynchronously record 'events', which are categorised as large shifts in light intensity. This is a novel representation of data that has overarching benefits that are yet to be fully explored in a plethora of applications. The many purported benefits of event-based cameras have the potential to revolutionise efficient, low-power computer vision due to their efficient encoding of information. Since the cameras are asynchronous in nature they allow for low-latency feeds with very little motion blur and other similar visual artefacts. As well as this their pixel structure allows for a very high dynamic range, meaning they can be used even when bright sunlight and dark shadows are encompassed in the same scene.

With the emergence of the Internet of Things (IoT), sensors have become, or at the very least will become, ubiquitous in everyday life. These devices operate using power at a premium since they need to perform adequately with a limited power supply (e.g., small battery cells etc.). This scarcity is further emphasised when attempting to do more power-intensive tasks. One such task is computer vision, which is becoming evermore prevalent as a means of human-computer interaction. Classical frame-based cameras are power intensive, and do not allow for a low active duty cycle (i.e., allowing for device sleep/idle time). The usual way in which this issue is alleviated is by making the system react to an event trigger by 'waking up' to work for a short period of time. Such events may include things such as; motion, timing, acceleration, or temperature. When compared to the function of event-based cameras, where event detection in built into the system, this can be seen as a stop-gap measure.

The high temporal resolution and dynamic range of event-based cameras also presents further benefits when compared to frame-based cameras. These features allow for videos to be represented in a very high fidelity format that preserves more data in fast-paced and brightly lit environments. This may lead to more reliable use of the data for computer vision and even other purposes where the exact environment the camera is set up in is uncertain and unpredictable.

1.2 Objectives

The objectives of this project were roughly divided into two main sections; main objectives and extensions. Objectives were created so as to allow for contingencies and fallbacks at every stage.

1.2.1 Main Objectives

- Understand structure of neuromorphic data and methods to preprocess it for inputting into an Artificial Neural Network (ANN).
- Evaluate different ANN models on neuromorphic data from external datasets.

  This allows for determining and comparing the performance of both traditional Neural Networks (NNs) as well of Spiking Neural Networks (SNNs). These networks can take two main forms:

  - Two networks sequentially processing data. The first would be reconstructing spiking data into an intensity video, and the second would be applying pre-existing computer vision networks to analyse the frame-based output.
  - One network that takes the spiking data as input to directly carry out the intended function.

  The system will initially be used to solve a classification task.

- Create practical set-up to experimentally record data.

  Using a neuromorphic camera, data can be experimentally captured to assess the performance of any built networks on more realistic unseen input data.

- Carry out Simultaneous Localisation and Mapping (SLAM) for a robot moving in an unknown trajectory.

  Use most efficient NN model to solve the more complex task of SLAM. The neuromorphic camera can be used to capture data from a room in the absence of large sets of labelled datasets.

1.2.2 Challenges & Contingency Plan

- Only using existing datasets rather than practical set-up.

  Since there is sufficient amounts of existing datasets for classification tasks, they can be split into training, validation, and testing sets themselves. This eliminates the need to generate more data for the testing process.

- Rather than focusing on SLAM the emphasis may be on object detection/recognition or gesture recognition

  The project has been segmented into individual milestones, and so even if the final milestone of a SLAM algorithm isn't achieved, it is easy to change the scope of the project to be a classification or gesture recognition network.

1.3 Report Structure

A brief outline of the report is given below:

Background (Chapter 2)

This chapter gives a background to the project subject. There are outlines of previous works in the field in order to highlight gaps in knowledge where more work can be done. It provides a good illustration of what issues there are and what value there would be in solving them.

Analysis and Design (Chapter 3)

Implementation (Chapter 4)

This chapter outlines some initial steps taken in the implementation process.

Testing and Results (Chapter 5)

Conclusion and Further Work (Chapter 7)

Chapter 2

Background

This chapter outlines background information required for understanding the basis for the project. The theory and literature serves to outline the main concepts used for neuromorphic data processing, as well as to reveal gaps in existing research that require solidifying.

2.1 Event Cameras

Event based cameras can be described as 'bio-inspired sensors that differ from conventional frame cameras: Instead of capturing images at a fixed rate, they asynchronously measure per-pixel brightness changes, and output a stream of events that encode the time, location and sign of the brightness changes' [1].

2.1.1 Benefits

Event-based cameras are purported to provide a number of benefits including;

- **High temporal resolution**

  The reason for this is that whereas frame-based cameras have a certain frame-rate, event-based cameras do not have this limitation, meaning the "blind time" between frames is eliminated. The reason for this is that the function of a frame-based camera is dependent on the global shutter to capture the light at a particular instant, whereas event-based cameras can be thought of as having individual shutters for each pixel that are shut whenever an event occurs.

- **High dynamic range**

  The reason for this is again the fact that each pixel has its own individual shutter, but as well as this they all use a logarithmic scale, meaning they function well from very bright to very dim environments as well as fast shifts between the two.

- **Low power consumption**

- **High pixel bandwidth**

  Each pixel can capture events at the rate of kHz. This has the effect of reducing blur since there is a very high temporal resolution to begin with. This makes the system very responsive and therefore ideal for real-time systems.

- **Efficient Encoding**

  Since events are asynchronous and spatially sparse (i.e there are mainly 0 values in the output matrix), the encoding is very efficient, as opposed to frame-based cameras that produce data that is very spatially dense.

The above benefits are very persuasive reasons to adopt neuromorphic cameras in many different applications. It is conceivable that if algorithms can make use of these benefits (since most classical algorithms play to the strengths of the data generated by frame-based cameras), real-time systems could be completely revolutionised.

2.1.2 Function

Event-based cameras differ from frame based cameras fundamentally, in that they do not rely on a global shutter closing at regular intervals to record information of a scene. Instead each pixel closes whenever it detects an 'event' occurring. The way such events are detected is dictated by the 'event generation model'[1].

Each pixel responds to changes in its log photo-current ($L = log(I)$, where I is the perceived brightness), giving the system a very high dynamic range. A recorded event 'k' has the format $e_k = (\mathbf{x}_k, t_k, p_k)$. This is known as the Address Event Representation (AER). The first value is the spacial location of the event ($\mathbf{x}_k = (x_k, y_k)^\top$), the second value t_k is the temporal location, and the final value $p_k \in 1, -1$ indicates the polarity of the event (i.e in which direction the brightness gradient was changing). The brightness increment between two events at the same pixel is given by the equation $\Delta L(\mathbf{x}_k, t_k) = L(\mathbf{x}_k, t_k) - L(\mathbf{x}_k, t_k - \Delta t_k)$. In a perfect (noise free) environment an event is fired whenever the brightness increment reaches a temporal contrast threshold. This relationship is shown in eq. (2.1).

$$\Delta L(\mathbf{x}_k, t_k) = p_k C (C > 0) \tag{2.1}$$

It should be noted that the value of C could be variable and therefore different for $p_k = \pm 1$. Additionally, we can approximate the temporal derivative of a pixels brightness by substituting eq. (2.1) into the Taylor expansion given in eq. (2.2). eq. (2.3) shows the resulting format for the temporal derivative.

$$\Delta L(\mathbf{x}_k, t_k) \approx \frac{\delta L}{\delta t}(\mathbf{x}_k, t_k)\Delta t_k \tag{2.2}$$

$$\frac{\delta L}{\delta t}(\mathbf{x}_k, t_k) \approx \frac{\Delta L(\mathbf{x}_k, t_k)}{\Delta t_k} = \frac{p_k C}{\Delta t_k} \tag{2.3}$$

It should however be noted that the approximation in eq. (2.2) is only true under the assumption that Δt_k is exceedingly small. Since unlike frame-based cameras we do not measure absolute brightness, this is an indirect way of measuring and keeping track of the brightness within the frame.

fig. 2.1 shows the basic functionality of an event-based camera. Shown in **(a)** is the simplified circuit diagram of the DAVIS pixel, which in **(b)** is used to convert light into events (shown in real life in images **(c)** and **(d)**). **(e)** shows how this setup would view a white square rotating on a black disk. It is a stream of events going from the past in green to the present in red. These events can then be seen overlaid on a natural scene in **(f)**.

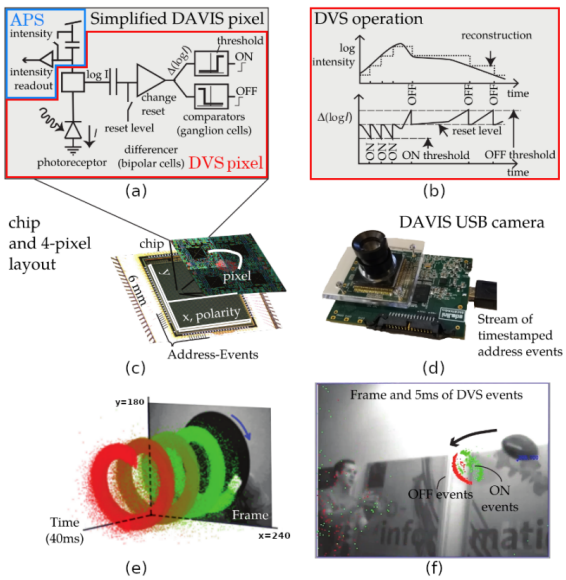


Figure 2.1: A summary of the functionality of the DAVIS event based camera[1].

2.2 Spiking Neural Networks and Neural Heterogeneity

Section 2.1.1 lists some persuasive reasons for utilising neuromorphic systems, but there still many challenges posed when attempting to do so. For example, each pixel only responds to brightness change, but the problem is that such a change could be a result of not only scene changes, but also the position of the camera within the scene. For this reason most neuromorphic systems have currently been limited to stationary cameras. As well as this, the system is especially prone to stochastic noise due to inherent shot noise in photons and from transistor circuit noise [1]. For this reason eq. (2.1) can only be said to be true under ideal conditions. A more realistic model for events fired is a probabilistic event generation model. These take into consideration the aforementioned sources of noise. One such model is given by P. Lichtsteiner *et al.*[6], where sensor variation measurements suggested a normal distribution centred around C for event triggers.

In order to tackle issues such as the ones due to noise, it is useful to look at existing examples of spiking neural systems, such as a biological brain. It is known that the brain is heterogeneous on every scale, in the past this was thought to be simply a by-product of noisy processes, but more recently it can be shown that by adding heterogeneity to Spiking Neural Networks (SNNs), a more stable and robust system can be created[7], indicating this heterogeneity has a more deep rooted purpose. As well as this the learned neural parameters tend to resemble what can observed experimentally. This may serve as an explanation for how the brain has evolved to deal with the many stochastic processing it encounters.

The foundations of SNNs are in computational neuroscience. The mechanisms of neurons in the brain are the inspiration behind creating ANNs with neurons that spike in the same way. Neurons in an a typical ANN have a weight, bias and activation function. This means that the output of the neurons can be summarised by eq. (2.4).

$$y = \theta(\sum_{j=1}^{n} w_j x_j - u_j) \tag{2.4}$$

This model was inspired by the biological neuron model shown in fig. 2.2. The function of the neuron is similar in that it produces an output based on a function of the input stream, but the form of this output is in the form of a 'spike' rather than a continuous function. The Σ shown in the diagram is actually the integration of all excitory and inhibitory input signals to the dendrites coming from the soma of the neuron. We can see that the electric potential of the neuron needs to exceed a certain threshold in order for the output of the neuron to be a spike. Spiking neural networks use a similar model of neurons in order to leverage the aforementioned benefits of neural heterogeneity.

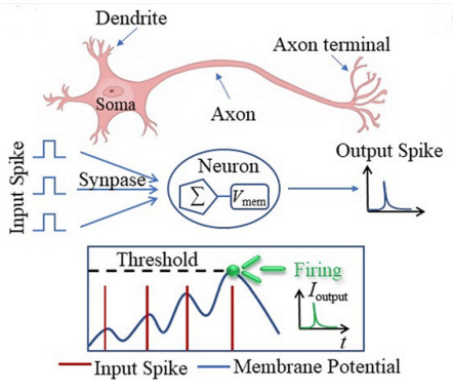


Figure 2.2: A diagram of structure and function of a biological neuron [2].

With SNNs it is theoretically possible to achieve very high energy efficiency, and when combined with a novel surrogate gradient and Recurrent Neural Network (RNN) as was experimented with in a paper by Bojian Yin *et al.*[8] excels in challenging tasks such as gesture recognition, and at some points even outperforms traditional ANNs. Alongside the aforementioned neural heterogeneity to combat stochastic processes we can see substantial improvements in previous SNN performance.

2.3 Existing Algorithms for Event Analysis

For SLAM, pose estimation and classification tasks the problem again is that classical systems heavily rely on the structure of conventional camera's outputs, and so there needs to be a radical paradigm shift in order to take events as inputs instead.

2.3.1 Optic-flow Methods

Since obtaining the equation for the temporal derivative in eq. (2.3), there is now a indirect measure of brightness and so a more classic computer vision techniques using optic-flow constraints can be utilised to characterise the events detected by pixels. In frame-based systems, optic flow methods create a flow-field that describe the displacement vector (signifying direction and magnitude of movement) for each pixel in the frame, and a similar derivation can be done for event data. A core constraint in this derivation is that the intensity of a local time-varying image region is constant under motion (for at least a short amount of time)[9]. eq. (2.5) is the resulting equations that shows the relation between the brightness gradient and the displacement of the pixel over a short period of time given its velocity[1]. It implies that if the motion is parallel to the edge, there is no event fired (since $v \cdot \nabla L = 0$) and conversely if the motion is perpendicular to the edge events are fired at their highest rate.

TODO: Write about how optic flow is often used in classification to see in a frame what the directions of motion are as well as the frame itself and the laplacian or edge map.

$$\Delta L \approx -\nabla L \cdot v \Delta t_k \tag{2.5}$$

2.3.2 Frame Integration of Event Streams

A common method of getting frame-like videos from event data is to use the event-to-frame integration method. One such method is outlined by *Wei Fang et al.*[10], which is the one used in the python package spikingjelly[11]. The method used is outlined in eq. (2.6).

Data in neuromorphic datasets are in the formulation of $E(x_i, y_i, t_i, p_i)$ that represent the event's coordinate, time and polarity. We split the event's number N into T slices with nearly the same number of events in each slice and integrate events to frames. Note that T is also the simulating time-step. Denote a two channels frame as $F(j)$ and a pixel at (p, x, y) as $F(j, p, x, y)$, the pixel value is integrated from the events data whose indices are between j_l and j_r:

$$j_l = \lfloor \frac{N}{T} \rfloor \cdot j$$

$$j_r = \begin{cases} \lfloor \frac{N}{T} \rfloor \cdot (j+1), & \text{if j} < \text{T - 1} \\ N, & \text{if j} = \text{T - 1} \end{cases}$$

$$F(j, p, x, y) = \sum_{i=j_l}^{j_r - 1} I_{p,x,y}(p_i, x_i, y_i) \tag{2.6}$$

where $\lfloor . \rfloor$ is the floor operation, $I_{p,x,y}(p_i, x_i, y_i)$ is an indicator function and it equals 1 only when $(p, x, y) = (p_i, x_i, y_i)$.

TODO: Write more about this method from the spikingjelly docs

2.4 Image Reconstruction Algorithms

Image reconstruction has been implemented for event data building on the direct optimised versions of Convolutional Neural Networks (CNNs). An example of this is the network named 'U-net'[12] which managed to reconstruct a video using 10M parameters to analyse events from an AER camera protocol. Recent work by Rebecq *et al.* illustrates a novel network architecture that reconstructs a video from a stream of events [3]. These methods are purported to allow the introduction of mainstream computer vision research to event cameras. Figure 2.3 shows an example of how converting spiking data to a video stream allows for use of classical computer vision algorithms.

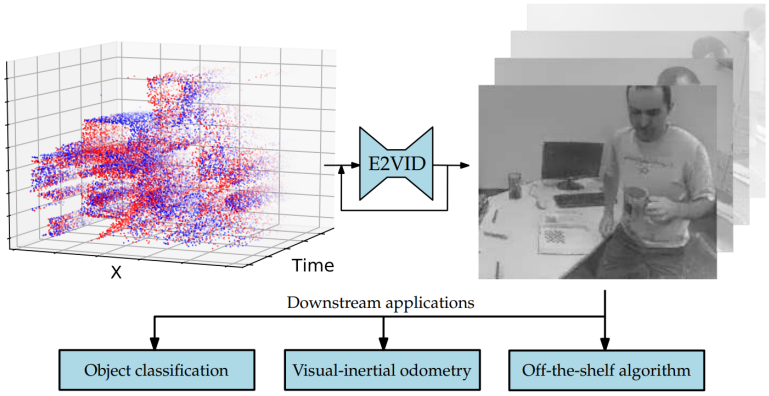


Figure 2.3: An illustration of the mapping of spiking data to video stream to apply off-the-shelf algorithms to[3].

A naive approach would be take each event $e_k = (\mathbf{x}_k, t_l, p_k)$ and assume that the firing was due to a brightness change above a threshold $\pm C$ which is a constant that could be set by the user. If this was the case events could be directly integrated to recover the intensity map of images. however, the value C in reality does not remain constant and is heavily dependent on other factors such as event rate, temperature, and sign of brightness change. The implementation outlined instead makes use of a Recurrent Neural Network (RNN), that takes as input sets of events within a spatio-temporal window. For example, a stream of events will be broken down into sequences given by $\epsilon_i \; \forall i \in [0, N-1]$. Since each sequence is of fixed length N the framerate of the output video from the RNN is proportional to the event rate. Figure 2.4 shows the functionality of such a network. Each event window ϵ_k is converted to a 3D event tensor and passed into the network along with the last K constructed images to generate the latest iteration of the image. It is clear from this that each new image is constructed by fusing the previous K images with the new stream of events.

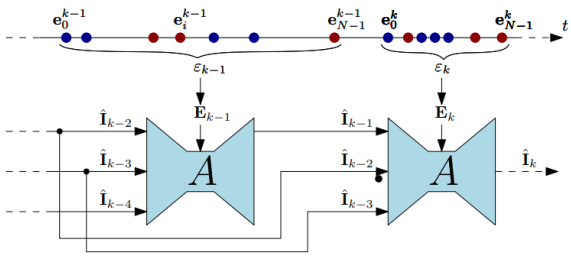


Figure 2.4: An overview of RNN used to generate video from sets of events[3].

2.4.1 Black-box Network

A method such as the one described in section 2.4 allows for the use of hugely researched and well documented computer vision algorithms for classification and other tasks while maintaining the benefits of event-based cameras. However, it may be possible to bypass the intermediate step of video reconstruction altogether, and simply create a model that simply acts as a black box and can be trained to give the required output directly from a neuromorphic input. Figure 2.5 shows how a frame-based video is converted to a continuous stream of events from which snapshots of events can be taken. It should be noted that the distribution of the data from the neuromorphic camera is much more dense than the frame-based video, which means that the motion blue visible in the video should not be a problem with the new representation (as explained in section 2.1.1).

In a paper by Arnon Amir *et al.*[4] a system was created to perform gesture recognition from event-based data. It makes use of the a system such as the one shown in fig. 2.5 to act as one of a set of temporal filters in a cascade. This cascade feeds into a convolution layer and in the end a winner takes all filter is applied to identify the gesture. The whole network can be seen in fig. 2.6.

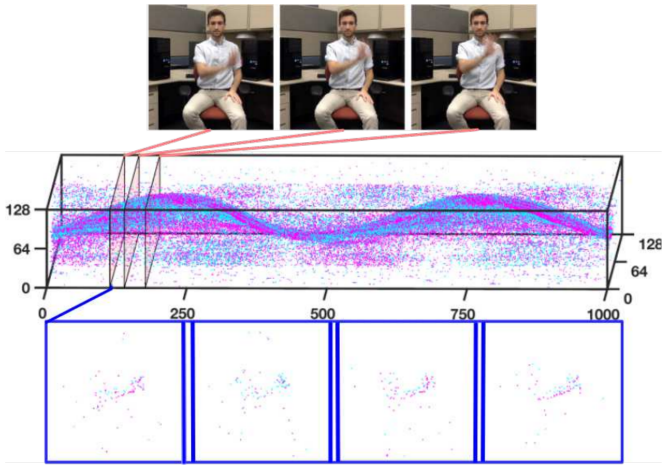


Figure 2.5: An illustration of a temporal filter that caches events fired from a camera[4].

The intermediate representations of all the layers can also be seen, showing how important features are being identified similar to how they would have been with traditional frame-based inputs.

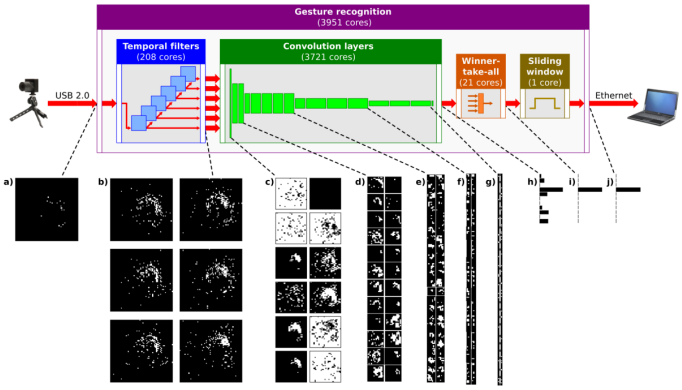


Figure 2.6: A network designed to perform gesture recognition using neuromorphic input by making use of cascading temporal filters[4].

There have been other implementations of systems to carry out complex tasks such as classification, and each has a different way of dealing with spiking input data other than with temporal filters. For example a paper by Xavier Lagorce *et al.*[13] moves towards building time surfaces from events to feed into a neural network, whereas Yin Bi *et al.*[14] propose using a non-uniform sampler to create a graph from events to then feed into a network of so-called graph convolution networks.

2.5 Temporally Aware Deep Learning and Classification Models

2.5.1 3-Dimensional Convolutional Neural Network

Convolutional neural networks have been shown to be much more effective when processing images that networks built solely with dense, fully-connected layers TODO: add references here . This is because they are able to better identify spacial patterns within an image as a kernel spans more than one pixel. For this reason the basic architecture was to have an input layer (the structure of which is dependent on the input format), followed by a series of hidden convolutional layers of varying parameters. However, since the inputs to systems handling event data are in fact 3D tensors of multiple images (i.e. the integrated frame video generated from the camera events, the process of which is described in section 2.3.2) a typical convolutional network is not sufficient

to capture the temporal patterns in the data. Typically 2D convolution layers can take as input images with three channels (usually RGB), and so feasibly this could be extended to more channels for each frame of the video, but this is not a scalable approach. Instead 3D convolutional layers can be used [15]. For these layers multiple frame time-slices are concatenated into a 3D tensor so that convolutions can occur over both the spacial and temporal dimensions. This allows for patterns to be detected across multiple adjacent frames.

2.5.2 Long-Short Term Memory Networks

When attempting to store information learned in previous frames of a multi-framed input to a network, recurrent back-propagation is a long process due to decaying error backflow. Long-short term memory (LSTM) networks[16] address these issues with efficient, gradient-based methods. These systems were tested to run much more quickly, efficiently and successfully than their recurrent network counterparts. They also allow networks to learn more complex patterns such as long-time-lag tasks that previously plagues ANNs.

2.5.3 Gated Recurrent Unit Networks

TODO: find papers on GRU

2.5.4 Convolutional LTSM Network

LSTM networks show promise in their ability to find patterns not only on an frame-by-frame basis but also in the time dimension. These spatio-temporal patterns are much more effective for classifying image sequences that just spatial patterns since information is carried throughout the frames to find overall movements and gestures. These LTSMs utilise fully connected layers to find patterns in each frame, however it has been shown that convolutional neural networks produce much better results when operating on image data, and so it would stand to reason that LTSMs would benefit from their structure as well. ConvLSTM, developed by *Xingjian SHI et al.*, showed great promise, and in their experiments captured spatio-temporal correlations better and more consistently FC-LSTM, outperforming it by a sizeable margin in the application of forecasting.

TODO: Write more about convLTSM networks etc.

2.6 Hardware and Software

2.6.1 Programming Languages

When choosing a programming language for the project there were a few choices that are most often chosen by developers; Python[17], R[18], and C++[19]. Python is the most popular choice due to the ease with which algorithms can be developed using it. Python features an extensive list of libraries and debugging functions that are invaluable when creating machine learning algorithms in particular, making the language of choice for this project. It is also important to note, however, the benefits of the other language options. C++ often results in programs with impressive performance due to the ability it grants to make low-level processes more efficient. Unfortunately this fine-level control also opens up programmers to more a demanding and time-intensive programming experience, with much more code writing and debugging to be done manually. R would also be a great choice for machine learning, and shares many similarities with Python, being open-source and having a huge community of developers constantly building libraries and tools. It has a different approach to machine learning, with a a more statistical analysis emphasis. Therefore Python remains the best choice for a more general approach for data processing.

2.6.2 Machine Learning Frameworks and Software

PyTorch

Pytorch[20] is a relatively new framework for machine learning, and provides a developer friendly way to write machine learning code. It is a more 'pythonic' approach to code abstraction that its competition in the space, and the *torch.nn.module* gives access to clear, reusable module definitions in an Object-Oriented Programming manner. It also allows for simple data parallelism, so that

batch processing can easily be split over different sets of hardware. It also has an intuitive debugging experience since itt can use standard debugging tools such as PyCharm and pdb.

Tensorflow

Tensorflow[21] is the older and more widely adopted machine learning library. It provides a more robust set of functionality with clear documentation. In terms of deployment it is the clear favourite as it allows models to be deployed on specialised servers and even on mobile. When visualising data software such as TensorBoard are ideal as it includes functionality to display model graphs, variables, histograms and much more. As well as this, Keras[22] is a framework developed by Google, and uses a primarily Tensorflow based back-end. It provides an easy to use API for fast prototyping abd high levels of abstraction. It is used commercially by a plethora of companies and has a vast and highly developed research community. For these reasons Keras using a Tensorflow back-end were chosen for this project.

2.6.3 Other Software

SpikingJelly

SpikingJelly[11] is an open-source deep learning framework for Spiking Neural Network (SNN) based on PyTorch. It allows for the processing of many often-used datasets in the neuromorphic community (Some of which are described in section 2.7), as well as a simple set of classes for building SNNs or converting ANNs to SNNs. The library, which was mainly co-developed by Multimedia Learning Group, Institute of Digital Media (NELVT), Peking University and Peng Cheng Laboratory, can be installed directly via the *pip* command. Using it, data can be loaded as event streams, as well as integrated frames (described in section 2.3.2) of varying frame lengths or frame-rates. As well as this the package features clear documentation and tutorials to begin analysing neuromorphic data.

Nengo

TODO: Write about nengo and brian etc.

2.6.4 Cloud Environments

Since Python is the language of choice for the project, Python Notebooks are a good choice for code segmentation and presentation. They allow for python code to be written in executable cells, so that their output as well as other text can be presented in a full document. This makes the code easy to understand for others, and good for development as well. Python notebooks can be run locally on a web server, as well as online on a cloud service. Many machine learning frameworks and algorithms make use of hardware acceleration using GPUs or TPUs. This means that code runs much faster on more powerful machines with this specialised hardware in them, which was not the case for hardware readily available during the course of the project. For this reason cloud services provided by the likes of Google and Amazon Web Services are a good alternative to physically owning hardware. They allow for the renting of GPUs etc. from their own servers, so that code can be run on them via their respective web services.

The two main web services available at this time are Google Colaboratory[23] and AWS Sagemaker Studio Lab[24]. In terms of hardware, both services offer access to great GPUs, though sagemaker offers the more powerful T4 GPU at the free tier. This benefit, however, is not entirely relevant as a pro subscription would be necessary with either service to make use high-RAM runtimes. Due to Colaboratory's better share-ability and wider adoption, it was chosen as the service for this project.

2.7 Existing Datasets

There already exists many repositories of recorded neuromorphic data to get familiar with spiking data. In this section there are some examples of such datasets and a quick overview of their contents.

2.7.1 Neuromorphic Datasets

The below datasets are created using one of a variety of event-based cameras available on the market. the function of each of the cameras is fundamentally similar (as described in section 2.1.2) but they also have some differences between them. The cameras widely available today are shown in fig. 2.7.

	Supplier	iniVation			Prophesee				Samsung			CelePixel		Insightness
	Camera model	DVS128	DAVIS240	DAVIS346	ATIS	Gen3 CD	Gen3 ATIS	Gen 4 CD	DVS-Gen2	DVS-Gen3	DVS-Gen4	CeleX-IV	CeleX-V	Rino 3
Sensor specifications	Year, Reference	2008 [2]	2014 [4]	2017	2011 [3]	2017 [67]	2017 [67]	2020 [68]	2017 [5]	2018 [69]	2020 [39]	2017 [70]	2019 [71]	2018 [72]
	Resolution (pixels)	128 × 128	240 × 180	346 × 260	304 × 240	640 × 480	480 × 360	1280 × 720	640 × 480	640 × 480	1280 × 960	768 × 640	1280 × 800	320 × 262
	Latency (µs)	12μs @ 1klux	12μs @ 1klux	20	3	40 - 200	40 - 200	20 - 150	65 - 410	50	150	10	8	125μs @ 10lux
	Dynamic range (dB)	120	120	120	143	> 120	> 120	> 124	90	90	100	90	120	> 100
	Min. contrast sensitivity (%)	17	11	14.3 - 22.5	13	12	12	11	9	15	20	30	10	15
	Power consumption (mW)	23	5 - 14	10 - 170	50 - 175	36 - 95	25 - 87	32 - 84	27 - 50	40	130	-	400	20-70
	Chip size (mm^2)	6.3 × 6	5 × 5	8 × 6	9.9 × 8.2	9.6 × 7.2	9.6 × 7.2	6.22 × 3.5	8 × 5.8	8 × 5.8	8.4 × 7.6	15.5 × 15.8	14.3 × 11.6	5.3 × 5.3
	Pixel size (µm^2)	40 × 40	18.5 × 18.5	18.5 × 18.5	30 × 30	15 × 15	20 × 20	4.86 × 4.86	9 × 9	9 × 9	4.95 × 4.95	18 × 18	9.8 × 9.8	13 × 13
	Fill factor (%)	8.1	22	22	20	25	20	> 77	11	12	22	8.5	8	22
	Supply voltage (V)	3.3	1.8 & 3.3	1.8 & 3.3	1.8 & 3.3	1.8	1.8	1.1 & 2.5	1.2 & 2.8	1.2 & 2.8		1.8 & 3.3	1.2 & 2.5	1.8 & 3.3
	Stationary noise (ev/pix/s) at 25C	0.05	0.1	0.1	-	0.1	0.1	0.1	0.03	0.03		0.15	0.2	0.1
	CMOS technology (nm)	350	180	180	180	180	180	90	90	90	65/28	180	65	180
		2P4M	1P6M MIM	1P6M MIM	1P6M	1P6M CIS	1P6M CIS	BI CIS	1P5M BSI			1P6M CIS	CIS	1P6M CIS
	Grayscale output	no	yes	yes	yes	no	yes	no	no	no	no	yes	yes	yes
	Grayscale dynamic range (dB)	NA	55	56.7	130	NA	> 100	NA	NA	NA	NA	90	120	50
	Max. frame rate (fps)	NA	35	40	NA	NA	NA	NA	NA	NA	NA	50	100	30
Camera	Max. Bandwidth (Meps)	1	12	12	-	66	66	1066	300	600	1200	200	140	20
	Interface	USB 2	USB 2	USB 3		USB 3	USB 3	USB 3	USB 2	USB 3	USB 3			USB 2
	IMU output	no	1 kHz	1 kHz	no	1 kHz	1 kHz	no	no	1 kHz	no	no	no	1 kHz

Figure 2.7: A table listing widely available event-based cameras and their respective features[1].

NMNIST

This dataset is a spiking version of the original frame-based MNIST dataset [25][5]. It is identical to the original MNIST dataset, which is a set of handwritten digits, in all ways (including scale, size and sample split) bar one - it was captured using an ATIS sensor mounted on a motorised pan-tilt unit. This sensor moved while viewing the MNIST examples on an external monitor.

For each item in the dataset there is a binary file which has a list of events. Each event is characterised by a 40 bit unsigned integer. The integer gives the following information of a particular event:

- bit 39 - 32: X location (in pixels)
- bit 31 - 24: Y location (in pixels)
- bit 23: Polarity (0 for OFF, 1 for ON)
- bit 22 - 0: Timestamp (in microseconds)

An example of a visualisation of this data is shown in fig. 2.8, where the image used from the MNIST dataset is on the right in part **(b)** and the resulting spikes from the event-based camera are shown on the left in **(a)**. Here, 'on events' are events where the intensity of a the particular pixel increased by an increment greater than a threshold, and the 'off events' are when the intensities decrease by a increment greater than a threshold. The representation is clearly very different to the one given by a classical camera, and therefore the use of such data has to have a different approach to classical techniques.

DVS128 Gesture

This dataset is a set of 11 hand gestures from 29 subjects under 3 illumination conditions. It was created to help create a real-time gesture recognition system that utilises the low power capabilities of event-based cameras[26].

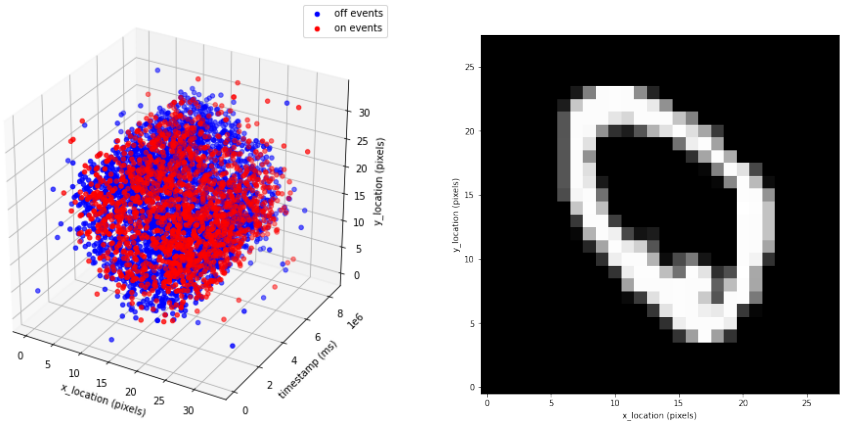
TODO: Write more about this dataset

CIFAR10-DVS

TODO: Write About The Event-Camera Dataset

Event-Camera Dataset

TODO: Write About The Event-Camera Dataset



(a) (b)

Figure 2.8: A visualisation of events, **(a)**, from a single training sample from the NMNIST dataset, **(b)**.

Heridelberg Spiking Datasets

The Spiking Heidelberg datasets for spiking neural networks[27] are useful for realising that spiking data is useful for so many more applications than computer vision. This dataset is split into two; The Spiking Heidelberg Digits (SHD) dataset and the Spiking Speech Command (SSC) dataset. Both of these datasets are audio-based classification datasets for which input spikes and output labels are provided.

2.7.2 Non-neuromorphic Datasets

Other data-sets such as fashion-MNIST could also be converted to spiking times by treating image intensities as input currents to model neurons, so that higher intensity pixels would lead to earlier spikes, and lower intensity to later spikes, as was done by Nicolas Perez-Nieves *et al.*[7]. This avoids having to own an event-based camera to create data (as was done with the NMNIST dataset mentioned in section 2.7.1), which is useful since the cameras are expensive and difficult to get a hold of in general.

2.8 Evaluation Metrics

The evaluation plan is as given by the typical machine learning pipeline[28]. TODO: Change this to actual used evaluation metrics and check it fits properly in backround

For a classification task, when we obtain the results from the test dataset (as shown in table 2.1) we can calculate a variety of evaluation metrics that give various insights on our final model.

Labels	Predictions
1	1
1	2
3	8
9	9
6	9
⋮	⋮

Table 2.1: A table showing an example of results when inputting test data from NMNIST dataset[5] into the final model.

2.8.1 Confusion matrix

Confusion matrices act as a visualisation of a systems performance. It shows possible true labels as well as possible predicted labels on either side, and filled in are the number of results that fit in each segment. In table 2.2 the confusion matrix for the NMNIST dataset is shown as an example. It should be noted that a similar confusion matrix should be created taking each class as positive, then each metric can be calculated by taking the averages (as shown in section 2.8.6). For each of the cells the number of matching records are stored to calculate each of the evaluation metrics. The table includes True Positives (TP), False Positives (FP), True Negatives (TN) and False Negatives (FN).

		Predicted Class			
		1	2	3	...
Actual Class	1	TP	FN	FN	...
	2	FP	TN	TN	...
	3	FP	TN	TN	...
	4	FP	TN	TN	...
	5	FP	TN	TN	...
	⋮	⋮	⋮	⋮	⋱

Table 2.2: a table showing one particular confusion matrix for NMNIST dataset[5] for class 1 as the positive class.

2.8.2 Accuracy

The accuracy of the system is the proportion of samples correctly classified.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Note: classification error can also be used and is defined as $1 - accuracy$.

2.8.3 Precision

Precision is the proportion of positively predicted samples identified correctly.

$$Precision = \frac{TP}{TP + FP}$$

It should be noted that a high precision may mean that there are many false positives.

2.8.4 Recall

Recall is the proportion of actual positives correctly classified.

$$Recall = \frac{TP}{TP + FN}$$

It should be noted that a high recall may mean a lot of positive samples may be missed.

2.8.5 F-measure/F-score

This is defined as the harmonic mean of precision and recall in order to get one number as an average measure of performance.

$$F_1 = \frac{2 \cdot precision \cdot recall}{precision + recall}$$

2.8.6 Micro and Macro Averaging

Macro-averaging involves taking an average on the class level. Metrics are calculated for each class and then averaged at the end. Micro-averaging involves taking an average on the item level (i.e., taking the average of each of TP, FP, TN and FN to get the averages metrics).

Chapter 3

Analysis and Design

3.1 Requirements Capture

The datasets used for the evaluation of the networks in this project were NMNIST (Section 2.7.1), DVS128 Gesture (Section 2.7.1) and CIFAR10-DVS (Section 2.7.1), since these are the most suited for training on classification tasks. As well as this their recording conditions and topics are varied, meaning an evaluation of the system can be more robust and reliable. The event camera dataset for slam (Section 2.7.1) are also extensive and include side by side frames from a traditional camera, and so were ideal to use for testing the event reconstruction algorithms.

TODO: Find a place for this

3.2 End-to-end Event Classification Models

3.2.1 Data Pre-processing

3.2.2 3-Dimensional Convolutional Neural Network

```
_________________________________________________________________
 Layer (type)                 Output Shape              Param #
=================================================================
 conv3d_7 (Conv3D)            (None, 8, 34, 34, 512)    14336

 activation_17 (Activation)   (None, 8, 34, 34, 512)    0

 max_pooling3d_10 (MaxPoolin   (None, 4, 17, 17, 512)   0
 g3D)

 conv3d_8 (Conv3D)            (None, 4, 17, 17, 256)    16384256

 activation_18 (Activation)   (None, 4, 17, 17, 256)    0

 max_pooling3d_11 (MaxPoolin   (None, 2, 8, 8, 256)     0
 g3D)

 conv3d_9 (Conv3D)            (None, 2, 8, 8, 128)      11239552

 activation_19 (Activation)   (None, 2, 8, 8, 128)      0

 max_pooling3d_12 (MaxPoolin   (None, 1, 4, 4, 128)     0
 g3D)

 flatten_3 (Flatten)          (None, 2048)              0

 dense_6 (Dense)              (None, 256)               524544

 activation_20 (Activation)   (None, 256)               0

 dense_7 (Dense)              (None, 10)                2570

 activation_21 (Activation)   (None, 10)                0
```

```
=================================================================
Total params: 28,165,258
Trainable params: 28,165,258
Non-trainable params: 0
_________________________________________________________________
```

Listing 3.1: Overview of layers in 3D convolutional network

A 3D convolutional network as described in section 2.5.1 was implemented to work directly on the event data. An overview of the layers present in the network, together with the shape of their outputs and number of trainable parameters can be seen in listing 3.1.

3.3 Convolutional LSTM Network

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv_lstm2d_6 (ConvLSTM2D)   (None, 8, 34, 34, 64)     150016

max_pooling3d_4 (MaxPooling  (None, 8, 17, 17, 64)     0
3D)

batch_normalization_6 (Batc  (None, 8, 17, 17, 64)     256
hNormalization)

conv_lstm2d_7 (ConvLSTM2D)   (None, 8, 17, 17, 32)     110720

max_pooling3d_5 (MaxPooling  (None, 8, 9, 9, 32)       0
3D)

batch_normalization_7 (Batc  (None, 8, 9, 9, 32)       128
hNormalization)

conv_lstm2d_8 (ConvLSTM2D)   (None, 9, 9, 16)          27712

max_pooling2d_2 (MaxPooling  (None, 5, 5, 16)          0
2D)

batch_normalization_8 (Batc  (None, 5, 5, 16)          64
hNormalization)

flatten_2 (Flatten)          (None, 400)               0

dense_4 (Dense)              (None, 256)               102656

dense_5 (Dense)              (None, 10)                2570

=================================================================
```

Listing 3.2: Overview of layers in Convolutional LTSM network

A convolutional LSTM network as described in section 2.5.4 was implemented to work directly on the event data. An overview of the layers present in the network, together with the shape of their outputs and number of trainable parameters can be seen in listing 3.2.

3.4 Two-phase Intensity Reconstruction Models

In order to make use of existing highly-tested and documented computer vision techniques, intensity reconstruction models were first applied to events.

Chapter 4

Implementation

4.1 Two-phase Intensity Reconstruction Models

The models described above allow for the system to learn directly on the spiking data. Another approach that was taken was to attempt to utilise video reconstruction networks to the event data so that more classical models and architectures could be used to patterns in the data.

4.1.1 Reconstruction Algorithms

E2VID

E2VID, as described in TODO: Reference background reading here , is a state-of the art network based on UNET that reconstructs intensity videos from events data. Having gotten the output from the network (which on test data had 90% accuracy TODO: check accuracy figure)

TODO: all of the following;

- Similar to denoising network CNN more easily able to find patterns in data
- gesture intensity maps equivalent to 100fps but retains high frequency information without too much noise
- spikingjelly for integrating frames which is widely used approach

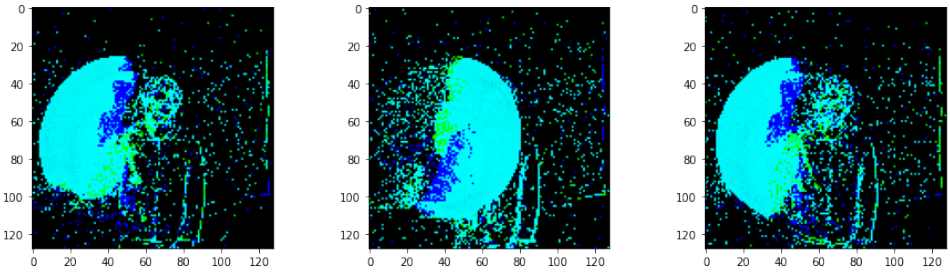
4.1.2 Classification Models on Reconstructed Video Output

4.2 End-to-end Event Classification Models

4.2.1 Data Preprocessing

A common method of processing events to get a stream of events is given in section 2.3.2. This technique was applied to three readily available data-sets; NMNIST, DVS128 Gesture and CIFAR10-DVS.

The intensity frames of a hand gesture in fig. 4.1 shows the motion of an arm moving in a clockwise direction. For each instance of a gesture the event stream was split into 20 frames. This made the processing of the data easy for some of the networks described later in the chapter, since the frames could be packed into 3D tensors of equal size. This means that the frames are not synchronous like they would be from an frame-based camera, and the amount of time represented by each frame is varying. As well as this, since the event streams are of varying length in the time dimension, some videos are reconstructed to a more granular scale than others. It is apparent that in this particular sample a relatively large amount of time is being compressed into every frame, resulting in a large amount of motion being visible. With more frames being created for each sample this problem would be alleviated and more easily distinguishable features may be seen. However if too many frames are taken, not only are processing times greatly increased, events may be sparse and not show any visible pattern in any one frame. A benefit of having this blurring



(a) (b) (c)

Figure 4.1: Three contiguous intensity frames (a, b and c), created from the DVS128 Gesture dataset with a person moving their right hand clockwise.

event in the intensity frames is that the direction and degree of motion can be seen in the frames, and so the networks can also pick up these patterns in their classification process.

TODO: Add more photos of integrated frames with more and less frames

TODO: Include image and explanations of frame integration

TODO: Add stuff about possible nlp-like networks.

Time-relative Event Segmentation

In order to begin analysing neuromorphic data, it was pre-processed it into a form that a NN can take as input. One such method of doing so was to segment the events into groups based on their timestamp. Figure 4.2 shows a visualisation of intensity maps created from the NMNIST[5] dataset. The set of all events was split into eight segments, where each segment included events within a range of 1×10^6 ms (i.e., $0 \rightarrow 1$, $1 \rightarrow 2$, ..., $7 \rightarrow 8$). This way the data representation shifted to somewhat get back to a set of frames that mimicked the video output usually seen from everyday cameras. **(a)** shows the segmented events visualised in three dimensions (x_location, y_location and timestamp). In **(b)** these events were projected onto the two dimensional plane (of x_location and y_location), then the plots for on events and off events are shown separately. Finally in **(c)** an intensity map was created from the projected events. Each pixel in the intensity map grid was initialised to 0, and for every on event 1 was added to the cell, and for every off event 1 was subtracted from the cell. It was clear that the resulting output greatly resembled the MNIST[25] sample recorded by the ATIS camera (As shown in fig. 2.8 in section 2.7).

Once a set of projections was made for each sample from the NMNIST dataset, it could be fed into a neural network in parallel. Ths could be done by simply flattening each intensity map and feeding all the cell values to a fully-connected dense layer in parallel, or the maps could be fed in as images to a convolutional layer. For the convolutional layer method a 3D tensor could be generated by stacking each of the intensity maps against each other and feeding them all directly into the first layer of the NN.

Neural Network Input Structure

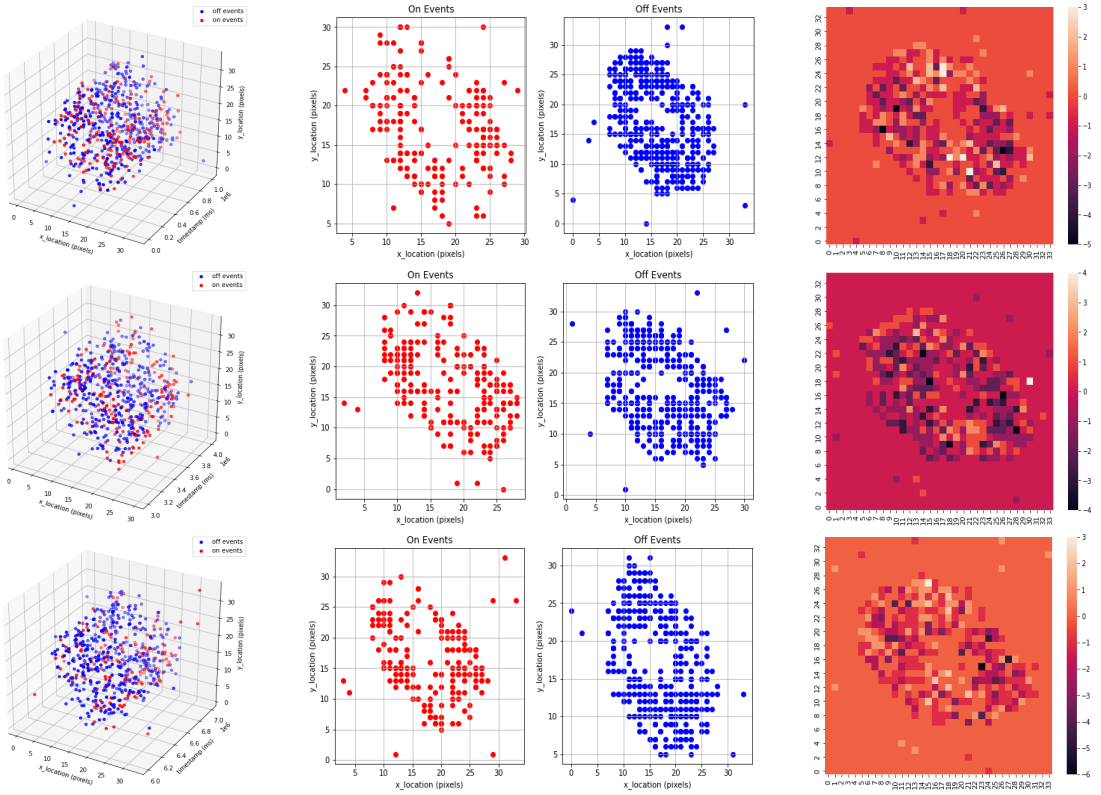
There were many ways with which a neural network could accept an input to the system. The ones considered for this project where;

- Flattening all the intensity maps from segmented event stream into a vector to be input to a single layer of neurons.
- Passing a 3D tensor to an convolutional layer as input.

4.2.2 Classification Models

Pure Convolution Network

The 3D convolutional neural network in section 2.5.1 was altered to have the outputs of the hidden layer fed into a dense layer with 10 neurons to classify the correct class (0-9). Activation functions need to be present in the network to prevent all the layers from becoming equivalent



(a) (b) (c)

Figure 4.2: A visualisation of intensity maps created by segmenting events into bins of size 1×10^6 ms.

to a single one TODO: add references here (linear regression model). In order to learn more complex patterns activation functions are a necessary aspect of creating an artificial neuron (See eq. (2.4) in section 2.2). The most commonly used activation function in deep networks (and image recognition in particular), and in this network as well, is ReLU. TODO: add references here . Finally, the output layer has a sigmoid activation function. This function compresses the output smoothly between the ranges of 0 and 1. this means each of the outputs from the neurons can be interpreted as a the probability of the input being any one of the 10 classes. This means we can simply take the highest probability as the predicted class from the network.

Hyper-parameter Tuning:

In order to choose the most appropriate parameters for the system, multiple tests were run varying each of the possible hyper-parameters. The tests conducted were to determine;

- The number of hidden layers.
- The number of neurons per layer.
- The size of convolution kernels.
- The introduction of some fully connected layers after convolutional layers.

The network in each case was trained for multiple epochs. The performance of a typical network can be seen in fig. 4.3, where the performance on the training set steadily improves as the network progresses through epochs. Accuracy is one of the metrics defined in section 2.8, and the loss for the given network is called categorical cross-entropy loss. The formula used to calculate the loss is given by: $L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{C} y_c^{(i)} log(\hat{y}_c^{(i)})$, where there are N samples and C classes. $y_c^{(i)}$ is 1 when the class is correctly predicted and 0 otherwise and $\hat{y}_c^{(i)}$ is the predicted probability of class c for data-point i.

It was clear, however, that these results may be misleading since they only represent the efficiency of the system when classifying values within the training data-set. However, when looking

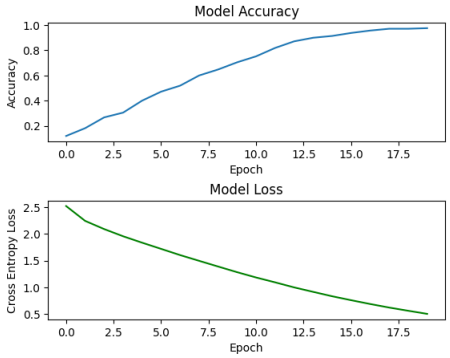
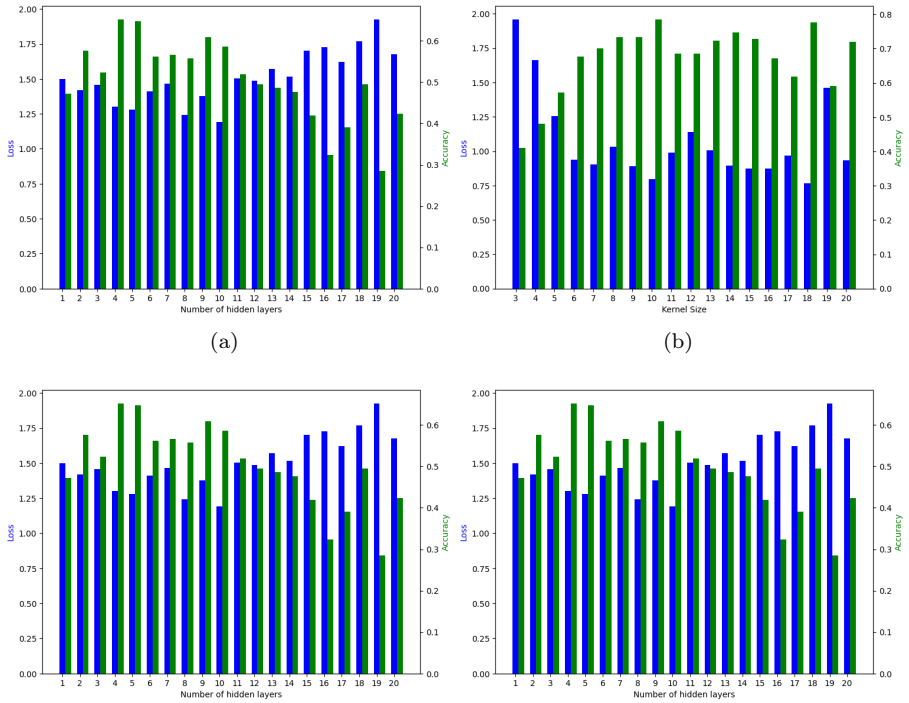


Figure 4.3: A figure showing classification accuracy and cross-entropy loss per epoch on training data for a typical network.

at the performance on an unseen test data-set, it is obvious that some of the features learnt do not easily translate to general trends in unseen data. This is known as over-fitting, and can be avoided by reducing the capacity of the data-set so that it does not learn information specific to the training set.

Model capacity is directly correlated to the n^o of filters, as well as the number of layers, and as the model recognises more patterns in the training data, so it is important to get the optimum value for the system. The size of the kernels has an effect on the scale of the information picked up by the system. With smaller kernels more local patterns are detected, whereas with larger kernels more global effects can be seen TODO: add image and reference here . As for the dense layers at the end of the network, it can be seen that better results were achieved as a result of it since global patterns can be further identified after the data has been processed by the convolution layers that have picked out the most important features.



(a) (b)

(c) (d)

Figure 4.4: A visualisation of intensity maps created by segmenting events into bins of size 1×10^6 ms.

Further testing was done with increasing kernel sizes, and with pooling layers added to the network. To find local patters filter kernels smaller than the image (32x32) are used. In earlier

layers the no of filters is large and the filter size is small to capture details. The no of filters decreases and the filter size increases later in the network for higher-level patterns. Pooling filters the image after convolution layers to pick out important features. TODO: add final system architecture here .

LTSM

Convolutional LTSM

TODO: Show model of custom Conv3D network

4.3 Spiking Neural Network

TODO: Cite nengo etc.

4.3.1 Synaptic Smoothing

4.3.2 Firing Rates

Post-training Scaling

Regularizing During Training

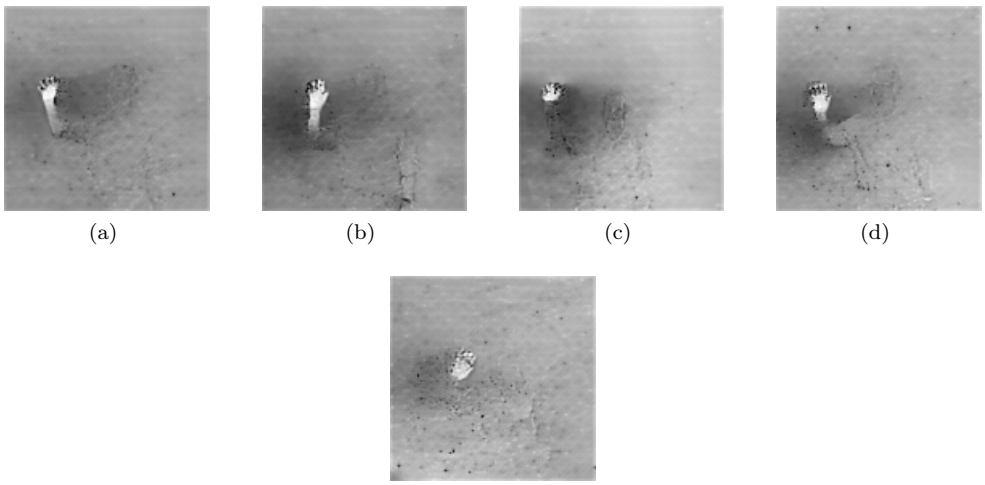
Chapter 5

Testing and Results

TODO: Write a nice intro to the chapter

It was interesting to note the performance differences between the previous networks and the described two-phase network.

5.1 Intensity Reconstruction



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

Figure 5.1: A waving motion being reconstructed from events captures by DVS128 event camera under different lighting conditions.The lighting conditions are as follows; **(a)** fluorescent led, **(b)** fluorescent, **(c)** lab lighting, **(d)** led lighting and **(e)** natural lighting.

TODO: Include image and explanations of reconstructions

5.1.1 NMNIST Dataset

5.1.2 DVS128 Gesture Dataset

5.2 Classification Results

Network	NMNIST	DVS128 Gesture
Conv3D Event Classifier	81.43%	47.22%
Conv2D LTSM Event Classifier	95.80%	
Custom LTSM Event Classifier	48.10%	
Conv2D LTSM Reconstruction Classifier	83.13%	19.26%

Table 5.1: A table showing classification accuracies of various models.

Chapter 6

Evaluation

6.1 Event Classification Networks

The benefits of the event driven camera (as described in section 2.1.1) were evident in the acquired results. TODO: Add images of integrated frames here . As opposed to tradition frame-based cameras, high frequency data is not lost when processing events. There are spikes for every event at a much more granular scale in the temporal dimension in the event camera when compared to the frame camera, meaning fast movements were captured more reliably since events are captured at the μs scale, no longer restricted by the frame-rates of modern cameras (which often results in motion blur).

It is, however, evident that modern computer vision techniques have been developed with frame-based cameras in mind, and so modern networks achieve good accuracy, and are able to find patterns well, on frame-based data. For this reason the common technique of integrating frame TODO: Add reference to spikingjelly integrating frames from background reading here results in regular frames, akin to the ones a regular camera generates, in order to feed into such networks. It does, however, still pose many benefits when compared to the frames from a regular camera. As mentioned previously information between frames is still not lost or degraded since the events are still captured and visible in each frame. As well as this, the frames generated from events inherently focussed on the points of interest in the image, since these were the only ones in motion in the frame. Most common architectures TODO: Add references to common gesture recognition architectures here for classical videos feature an intermediate layer to remove backgrounds and other noise from images from the image to focus on the points of interest. In this way the intermediate layer could be omitted when operating on the integrated event frames, since the output was already similar to an edge map TODO: Add image comparison here . It is conceivable, however, that in noisy environments with lots of motion this stage would still be necessary.

6.2 Two-Phase Video Reconstruction and Classification Networks

6.2.1 Video Reconstruction

The E2VID reconstruction network (as described in section 2.4) was used to recreate intensity videos from events. It was evident that the reconstructions created were robust and relatively to to life. The reconstructions were not effected by adverse lighting effects or fast motions TODO: Add image examples . This meant that modern computer vision techniques could still be applied to event data, while still preserving the many benefits the event model presents. It was interesting to note the performance of the reconstruction model on inputs with few moving parts. Since events are only triggered when there is an intensity change on any given pixel on the sensor, only regions with motion in them showed up as events, and the nature of all the space with no motion was not easily inferable. This can clearly be seen in fig. 5.1, where only parts of the scene were accurately reconstructed. This did not, however, pose much of a problem for tasks such as gesture recognition, since the motion is exactly what is being classified, however for other tasks such as object recognition it had to be ensured that there was some sort of motion of either the object or

the camera for the reconstruction algorithm to be effective.

Figure 5.1 also shows that event-cameras do indeed allow for higher fidelity video capture in a wider range of lighting conditions that frame-based cameras (as explained in section 2.1.1). In all lighting conditions the video reconstruction was largely the same, since the events triggered were very similar in all cases. The logarithmic characteristics of the event-sensor pixels are the reason for this, since the thresholds for the triggering of events is not static. Since the reconstructions are consistent across lighting conditions, this also means that the reconstruction is more reliable, and the classification algorithm works better in adverse conditions in general TODO: get figures and images to prove this .

6.3 Comparisons and Evaluations of Models

TODO: Write out the following;

- for NMNIST the small size of images means that reconstructions do no vary much from the rudimentary intensity maps (could be used as an illustration of how reconstruction works).
- Similar to denoising network CNN more easily able to find patterns in data
- gestures are more complicated to characterise, since the networks for NMIST can easily learn all possible features. As well as this the image size for NMNIST is very small
- intensity maps can be of much higher framerate since events are taken asynchronously
- gesture intensity maps equivalent to 100fps but retains high frequency information without too much noise
- most gesture recognition datasets involve removing backgrounds and focussing on the hand, our network automatically only sees moving objects
- gesture recognition robust for even harsh lighting conditions

Chapter 7

Conclusion and Further Work

Appendix A

First Appendix

Bibliography

[1] G. Gallego, T. Delbruck, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis *et al.*, "Event-based vision: A survey," *arXiv preprint arXiv:1904.08405*, 2019.

[2] X. Liang, X. Zhang, J. Xia, M. Ezawa, Y. Zhao, G. Zhao, and Y. Zhou, "A spiking neuron constructed by the skyrmion-based spin torque nano-oscillator," *Applied Physics Letters*, vol. 116, no. 12, p. 122402, 2020.

[3] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "Events-to-video: Bringing modern computer vision to event cameras," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 3857–3866.

[4] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza *et al.*, "A low power, fully event-based gesture recognition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7243–7252.

[5] G. Orchard, A. Jayawant, G. K. Cohen, and N. Thakor, "Converting static image datasets to spiking neuromorphic datasets using saccades," *Frontiers in neuroscience*, vol. 9, p. 437, 2015.

[6] P. Lichtsteiner, C. Posch, and T. Delbruck, "A 128 $\times$ 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.

[7] N. Perez-Nieves, V. C. Leung, P. L. Dragotti, and D. F. Goodman, "Neural heterogeneity promotes robust learning," *bioRxiv*, pp. 2020–12, 2021.

[8] B. Yin, F. Corradi, and S. M. Bohté, "Accurate and efficient time-domain classification with adaptive spiking recurrent neural networks," *arXiv preprint arXiv:2103.12593*, 2021.

[9] G. Gallego, C. Forster, E. Mueggler, and D. Scaramuzza, "Event-based camera pose tracking using a generative event model," *arXiv preprint arXiv:1510.01972*, 2015.

[10] W. Fang, Z. Yu, Y. Chen, T. Masquelier, T. Huang, and Y. Tian, "Incorporating learnable membrane time constant to enhance learning of spiking neural networks," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2021, pp. 2661–2671.

[11] W. Fang, Y. Chen, J. Ding, D. Chen, Z. Yu, H. Zhou, Y. Tian, and other contributors, "Spikingjelly," https://github.com/fangwei123456/spikingjelly, 2020, accessed: 2022-05-20.

[12] O. Ronneberger, P. Fischer, and T. Brox, "U-net: Convolutional networks for biomedical image segmentation," in *International Conference on Medical image computing and computer-assisted intervention*. Springer, 2015, pp. 234–241.

[13] X. Lagorce, G. Orchard, F. Galluppi, B. E. Shi, and R. 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, 2016.

[14] Y. Bi, A. Chadha, A. Abbas, E. Bourtsoulatze, and Y. Andreopoulos, "Graph-based object classification for neuromorphic vision sensing," in *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 2019, pp. 491–501.

[15] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[16] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[17] G. Van Rossum and F. L. Drake, *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace, 2009.

[18] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2021. [Online]. Available: https://www.R-project.org/

[19] B. Stroustrup, *The C++ programming language*, 3rd ed. Addison-Wesley, 1997.

[20] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, A. Desmaison, A. Kopf, E. Yang, Z. DeVito, M. Raison, A. Tejani, S. Chilamkurthy, B. Steiner, L. Fang, J. Bai, and S. Chintala, "Pytorch: An imperative style, high-performance deep learning library," in *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. [Online]. Available: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

[21] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M. Schuster, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng, "TensorFlow: Large-scale machine learning on heterogeneous systems," 2015, software available from tensorflow.org. [Online]. Available: https://www.tensorflow.org/

[22] F. Chollet *et al.* (2015) Keras. [Online]. Available: https://github.com/fchollet/keras

[23] Google, "Google colaboratory," https://colab.research.google.com/, 2017, accessed: 2022-05-20.

[24] Amazon, "Aws sagemaker," https://aws.amazon.com/sagemaker/, 2017, accessed: 2022-05-20.

[25] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[26] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. Di Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza *et al.*, "A low power, fully event-based gesture recognition system," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 7243–7252.

[27] B. Cramer, Y. Stradmann, J. Schemmel, and F. Zenke, "The heidelberg spiking data sets for the systematic evaluation of spiking neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[28] M. R. Antoine Cully and J. Wang, "Introduction to machine learning (co395), lecture 3," University Lecture, 2020.