Kingston University London
Faculty of Science, Engineering and Computing
Department of Computing

Neuromorphic Vision-Based Tactile Sensor for Robotic Grasp

Fariborz Baghaei Naeini

Supervisory Team:
Prof Dimitrios Makris
Dr Yahya Zweiri

Submitted in part fulfilment of the requirements for the degree of
Doctor of Philosophy in Artificial Intelligence
November 2020

Abstract

Tactile sensors are developed to mimic human sense of touch in robotics. The touch sense is essential for machines to interact with environment. Several approaches have been studied to obtain rich information from the contact point to correct robot's actions and acquire further information about the objects. Vision-based tactile sensors aim to extract tactile information by observing the contact point between the robot's hand and environment and applying computer vision algorithms.

In this thesis, a novel class of vision-based tactile sensors is proposed, "Neuromorphic Vision-Based Tactile Sensor" to estimate the contact force and classify materials in a grasp. This novel approach utilises a neuromorphic vision sensor to capture intensity changes (events) in the contact point. The triggered events represent changes in the contact force at each pixel in microseconds. The proposed sensor has a high temporal resolution and dynamic range which are suitable for high-speed robotic applications.

Initially, a general framework is demonstrated to show the sensor operations. Furthermore, the relationship between events and the contact force is presented. Afterwards, methods based on Time-Delay Neural Networks (TDNN), Gaussian Process (GP) and Deep Neural Networks (DNN) are developed to estimate the contact force and classify objects material from the accumulation of events. The results indicate a low mean squared error of 0.17N against a force sensor for the force estimation using TDNN. Moreover, the objects materials are classified with 79.12% accuracy which is 30% higher compared to piezoresistive force sensors.

This is followed by an approach to preserve spatio-temporal information during the learning process. Therefore, the triggered events are framed (event-frames) within a time window to preserve spatial information. Afterwards, multiple types of Long Short-Term Memory (LSTM) networks with convolutional layers are developed to estimate the contact force for objects with different size. The results are validated against a force sensor and achieve a mean squared error of less than 0.1N.

Finally, algorithmic augmentation techniques are investigated to improve the networks accuracy for a wider range of force. Image-based and time-series augmentation methods are developed to generate artificial samples for training the network. A novel time-domain approach Temporal Event Shifting (TES) is proposed to augment events by preserving the spatial information of events. The results are validated on real experiments which indicate that time-domain and hybrid augmentation methods improve the networks' accuracy significantly considering an object with a different size.

Acknowledgements

First of all, I would like to acknowledge my supervisors Prof. Dimitrios Makris and Dr. Yahya Zweiri for their immense support through this journey. Completion of this degree would not be possible without their guidance, knowledge, inspirational ideas and expertise on research. Beyond my research, both of my supervisors have had an incredible impact on the way to approach challenges in life with an infinite mindset.

I would like to express my gratitude to all academics in Science, Engineering and Mathematics faculty particularly Prof. Jean-Christopher Nebel, members of Digital Information Research Center and Game Theory group. Deep discussions in research seminars and meetings have improved my research skills in various fields. In addition, I would like to appreciate my colleagues in the computer science laboratory (SB1018) who continuously motivate and encourage me through this journey. Also, I would like to extend my sincere thanks to all my colleagues at Ipsotek Ltd for their encouragement and support.

A special thanks to Kingston University for providing the fund for this research as well as excellent equipment in laboratories, immense student support and all other facilities. Also, I would like to thank Khalifa University especially all members of staff at Khalifa University Center for Autonomous Robotic Systems for data collection which are employed in this thesis.

Finally, I would like to express my sincere gratitude to my parents, brother, sister-in-law and all the family members who supported me in all aspects my life. This accomplishment would not have been possible without their continuous encouragement and support throughout years of study.

Declaration

I, Fariborz Baghaei Naeini, declare that this thesis with title of "Neuromorphic Vision-Based Tactile Sensor for Robotic Grasp" contains no material that has been submitted previously, in whole or in part, for the award of any other academic degree or diploma. Except where otherwise indicated, this thesis is my own work". I confirm that:

- This work was done wholly while in candidature for a research degree at Kingston university.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at Kingston university or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my work.
- I have acknowledged all main sources of help.

‘I get very excited when we discover a way of making neural networks better and when that’s closely related to how the brain works.’
Geoffrey Hinton

Contents

List of Tables

List of Figures

List of Abbreviations

ARD Automatic Relevance Determination

CNN Convolutional Neural Network

ConvLSTM Convolutional Long Short-Term Memory

DCGAN Deep Convolutional Generative Adversarial Networks

DNN Deep Neural Network

DTW Dynamic Time Warping

DVS Dynamic Vision Sensor

F/T Force/Torque

FBG Fiber Bragg Grating

FEM Finite Element Model

FS Frame Shifting

GAN Generative Adversarial Networks

GPU Graphical Processing Unit

GP Gaussian Process

GRU Gated Recurrent Units

IFEM Inverse Finite Element Model

LED Light-Emitting Diode

LSTM Long Short-Term Memory

MAE Mean Absolute Error

MSE Mean Squared Error

PDMS PolyDiMethylSiloxane

PVDF PolyVinyliDene Fluoride

RMSE Root Mean Squared Error

RNN Recurrent Neural Network

SCG Scaled Conjugate Gradient

SNN Spiking Neural Networks

SVM Support Vector Machine

TDNN Time Delay Neural Networks

TES Temporal Event Shifting

UAV Unmanned Aerials Vehicles

VBM Vision-Based Measurement

VGG Visual Geometry Group

Chapter 1

Introduction

1.1 Introduction

This chapter presents a concise overview of the research problem investigated in this thesis. Section 1.2 identifies the challenges that motivated the work in this thesis. In section 1.3, the hypotheses of this thesis are defined and research questions are highlighted. In section 1.4, the primary questions of this research are defined which is followed by aim and objectives of this thesis in section 1.5. The main contributions of this study are presented in section 1.6. An overview of the methodology and achievements are provided in section 1.7 and 1.8 respectively. Finally, the structure of this thesis is stated in section 1.9.

1.2 Research Motivation

Human's dexterous hands are capable of grasping various objects successfully without a prior knowledge of shape, weight and friction coefficients. Skin receptors in the human fingers acquire information of object slippage to apply minimum force during precision gripping [1, 2]. Capability of the human touch sense inspires researchers to create artificial skins and various tactile sensors to enhance robots proficiency in human-robot interactions [3], manipulation tasks [4] and minimally invasive surgical applications [5, 6].

Robotic grasping techniques are divided into two main categories, power grasp and precision grasp, where the former focuses on stability and the latter requires high sensitivity and precision [7]. Simple tasks for taking and placing specific tools in structured environments can be performed through power grasps. However, delicate manipulation for surgical applications in

unstructured environments is required; therefore, the precision of the grasp method is essential to minimise the risk of grasping failure. Moreover, the extra applied force could result in object deformation or destruction. Therefore, grasping nonrigid or breakable objects may require a more complex grasping model. Some of the power grasping methods and contact models are reviewed in [8], which indicates significant progress in structured grasping applications. Recent grasping techniques have simplified the force regulations by using tactile and position sensors considering unstructured environments with high uncertainty [9–11].

Detection and measurement of crucial parameters such as magnitude and direction of the force, local strain and prediction of incipient slip based on the contact area facilitate grasping applications to perform adaptive regardless of object characteristics. To achieve acceptable and stable precision grasp, it is necessary to acquire contact area properties with low latency to feedback force into controllers and plan best grasping trajectories. Three main problems are investigated in this research as the following:

(i) **Force Estimation**: Measuring the contact force magnitude between the objects and gripper will assist controllers in order to correct the applying force to the object to prevent any failures such as destruction of breakable objects, deformation of soft objects and stability of the grasp. Moreover, acquiring magnitude of the force in three dimensions (force vectors) and mapping the force distribution on the grippers' surface will provide precise information about the object position and grasp state for multidimensional grippers.

(ii) **Material Classification**: Recognition of objects' materials is one of the challenging tasks in manipulations which assists the robot to adopt strategies for picking and placing objects. Different materials are varied in Young's modulus (elasticity properties) which lead to follow specific patterns in a grasp.

(iii) **Data Scarcity**: Machine learning methods are effective but require a large number of actual data. This inherent limitation prevented the wide adoption of machine learning in industrial automation because collecting and assembling large-scale datasets for robotic grasping and manipulation is challenging and expensive.

1.3 Hypotheses

Conventional cameras capture the absolute value of intensity in the scene synchronously. Typically, cameras have a high resolution and sampling rate of 30FPS with a high power-consumption. In contrast, neuromorphic vision sensors record intensity changes within few microseconds at

each pixel asynchronously. The dynamic behaviour of neuromorphic vision sensors allows the system to reduce the latency and power-consumption significantly. In this research, a neuromorphic vision sensor is employed to investigate the following hypotheses for the tactile sensing applications:

1. **Force Estimation**: The applied force to the object deforms the flexible membrane which changes the intensity in the contact area. The contact force can be estimated based on the intensity changes by using neuromorphic vision sensors.
2. **Material Classification**: Computational analysis of the intensity changes during a grasp can provide information about the properties of the object like material type. Since materials have a different Young's modulus, triggered events will have distinguishable pattern during a grasp.
3. **Data Scarcity**: Image-based and time-domain augmentation methods facilitate the data collection process in order to achieve high accuracy with a limited number of experiments.

1.4 Research Questions

The research questions of this thesis are as follows:

1. How a neuromorphic vision sensor can be used to acquire tactile information?
2. What are suitable machine learning techniques for contact force estimation and material classification, applied on data derived by neuromorphic vision sensor?
3. How synthetic data can be generated to improve machine learning accuracy for neuromorphic vision-based tactile measurement systems?

1.5 Aim and Objectives

This research aims to develop a novel class of vision-based tactile sensors using neuromorphic vision sensors. Utilisation of Dynamic Vision Sensors (DVS) for tactile sensing applications has the potential for lower latency, lower power consumption, and higher sensitivity compared to conventional vision-based sensors. In order to investigate the hypotheses stated above, the following objectives have been formulated:

1. Investigate machine learning approaches to estimate the contact force magnitude during a grasp from intensity changes captured by the neuromorphic vision sensors.

2. Development of a novel event-based material classification for objects with the same shape and size in a single grasp using machine learning techniques.

3. Investigate algorithmic methods to generate synthetic data to improve the models' accuracy.

1.6 Contributions

After a thorough literature review and ongoing study of relevant new publications the contributions of this project so far to the field are as follows:

1. The first neuromorphic vision-based sensor to measure the contact force and classify materials in a grasp using a neuromorphic camera (DVS).

2. A Time Delay Neural Network (TDNN) and a Gaussian Process (GP) to find the correlation between the triggered events and the contact force. A Deep Neural Network (DNN) to classify objects' material in a grasp.

3. A deep Long Short-Term Memory (LSTM) network with convolutional layers to estimate the contact force from spatio-temporal features.

4. Image-based and time-domain augmentation techniques applied on spatio-temporal event data to enhance the force estimation accuracy for an object with different size.

1.7 Methodology

In a robotic grasp, the object contact area with the fingertip changes the intensity level of the pixels (observation). The intensity changes are captured by a neuromorphic camera with high temporal resolution. Then, a machine learning algorithm is trained to correlate both spatial and temporal features of the intensity changes to the contact force (learn). In the learning process, the system observes a variety of experiments with different objects, range of force and speed to create a function which maps the intensity changes to the contact force. The output of the sensor provides the contact force which is crucial in many robotic applications such as pick and place of fragile objects (sense). Figure 1.1 demonstrates the framework for the proposed sensor.

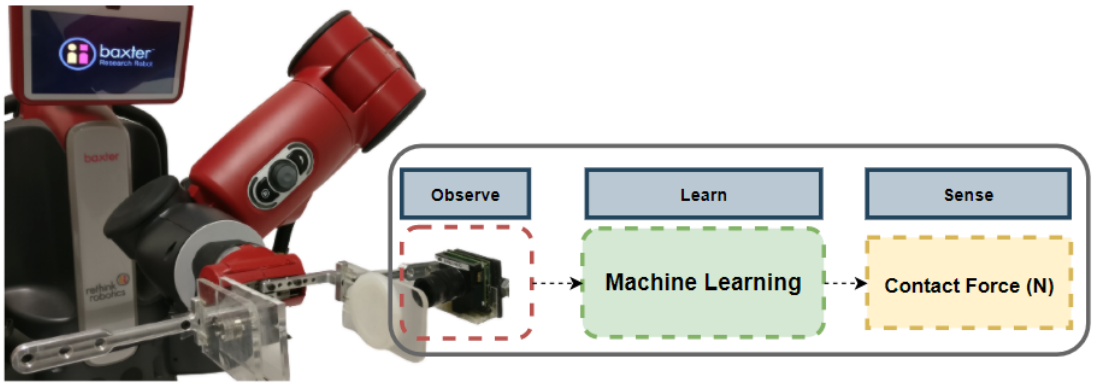


Figure 1.1: Diagram of the proposed novel sensor for the contact force estimation.

It should be noted that the design of experiments in this thesis in sections 3.6, , 4.5, 5.4 is mine, while their execution was performed by research team at Khalifa University Center for Autonomous Robotic Systems in a collaborative manner.

1.8 Achievements

List of peer-reviewed publications as part of this PhD thesis:

- Baghaei Naeini, F., AlAli, A.M., Al-Husari, R., Rigi, A., Al-Sharman, M.K., Makris, D. and Zweiri, Y., 2020. A novel dynamic-vision-based approach for tactile sensing applications. *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 1881-1893, doi: 10.1109/TIM.2019.2919354.
- Baghaei Naeini, F., Makris, D., Gan, D. and Zweiri, Y., 2020. Dynamic-vision-based force measurements using convolutional recurrent neural networks. *Sensors*, vol. 69, no. 16, 4469, doi:10.3390/s20164469.

1.9 Structure of the Thesis

The rest of this report is structured as follows: State-of-the-art work is reviewed in chapter 2 considering the research fields of tactile sensors, neuromorphic vision sensors and machine learning. In chapter 3, a dynamic-vision-based sensor is proposed to estimate the contact area and classify objects' material based on event temporal information. Afterwards, spatio-temporal deep learning models are investigated in chapter 4 to estimate the contact force. In chapter 5, augmentation techniques are proposed to generate synthetic training data to improve the

accuracy of the estimated force. Finally, findings of this thesis are summarised and future work is presented in chapter 6.

Chapter 2

Literature Review

This chapter reviews previous work related to the proposed novel vision-based tactile sensor. In section 2.1, various types of tactile sensors such as piezoresistive, capacitive and optical sensors are reviewed. Neuromorphic vision sensors and their applications are studied in section 2.2. Afterwards, relevant machine learning techniques are presented and discussed in section 2.3. Finally, the open research problems that drive this thesis are highlighted in section 2.4. The following diagram shown in Figure 2.1 demonstrates an overview of this chapter while the red blocks correspond to sections in this chapter.

2.1 Tactile Sensors

Tactile sensors are developed to acquire physical properties of the contact area via interaction with the environment. Similar to human skin receptors, tactile sensors are capable of measuring physical properties such as position, force, texture, stiffness, torque and temperature in the contact point. However, industrial applications like human-robot interaction [12], soft robotics [13] and object recognition [14] require different specifications for the sensor in terms of resolution, latency and accuracy. Thus, researchers develop a variety of measurements techniques to solve sensing challenges and overcome limitations of the sensors for diverse real-world applications which are reviewed in [15–18] comperhensively. In this chapter, an overview of popular non optical tactile sensing is provided, followed by a comprehensive review of optical tactile sensors with a focus on vision-based sensors.

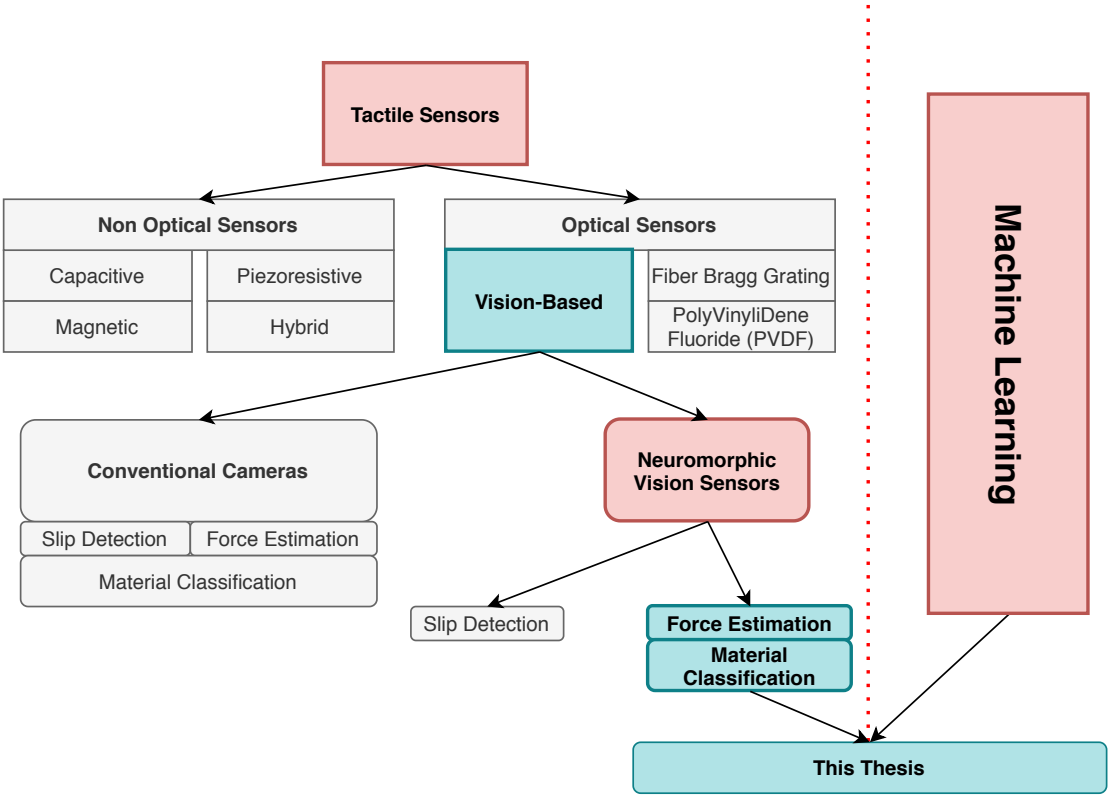


Figure 2.1: Diagram of related work in tactile sensing applications. The blocks in red presents sections of this chapter while blue blocks illustrates the primary focuses of this thesis.

2.1.1 Non Optical Tactile Sensors

Capacitive tactile sensors are well-known category that considers the variation of capacitance by changing load forces. These sensors comprise parallel conductive plates with a dielectric in between which can measure the applied force on the plates by decreasing distance between plates, and therefore, increasing the capacitance. In [19], a capacitive tactile sensor array is developed to recognize the texture pattern of the contact area. In another application, a capacitive sensor is designed with four capacitors to measure the horizontal and vertical forces [20]. Although the sensor can measure the displacement in different directions, electronic interference and sensor hysteresis have remained unsolved.

Piezoresistive tactile sensors consist of a conductive or semi-conductive elastomer that deforms under pressure. This deformation results in a change of material resistance that can be measured to identify touch and estimate the applied force [21, 22]. Furthermore, authors in [23] proposed an application of piezoresistive sensors to detect initial slip considering high-frequency components of the output signal. An application of object's localisation and orientation estimation is demonstrated in [24] using piezoresistive force sensors. The sensor is capable of localising the objects with high accuracy while the estimation of the object orientation has a poor resolution.

Another piezoresistive sensor with high durability and low hysteresis is developed in [25] to measure pressure under cyclic loading.

A magnetic tactile sensor is proposed in [26] with the capability of slip detection as well as estimation of the contact force in three dimensions. Some other approaches are utilising the tactile sensors to extract further information about the properties of the object and classify the material. For instance, a piezoelectric multi-functional sensor is used to acquire object hardness by rolling over the sensor on the surfaces of the objects [27].

A hybrid sensor in [28] composes of piezoelectric transducers, force sensor and inclinometer in order to classify six different materials. An artificial finger with embedded PolyVinyliDene Fluoride (PVDF) membrane and strain gauge sensors are used to classify various materials [29]. Another system is presented in [30], where two piezoresistive tactile sensors are utilised to classify softness of vegetable using a decision-tree machine learning technique. Other applications of object classifications using different types of force sensors and traditional machine learning algorithms such as Support Vector Machines (SVM) are presented in [31, 32].

However, most of the aforementioned tactile sensors have limited resolution, considerable hysteresis and interruptible to the electromagnetic disturbances. Furthermore, the sensors require a direct contact with the objects which limits the sensors applications.

2.1.2 Optical Tactile Sensors

This thesis focuses on optical tactile sensing techniques including camera-based methods that provide higher resolution, low hysteresis and resistant to electromagnetic disturbances. Optical sensors with transparent elastomer or rubber, wave emitters and receivers have been developed for precise tactile sensing applications [33, 34]. Wave emitters scatter the light to the surface of the elastomer while receivers capture the back-scattered beams from the surface. Study of the reflected beams regarding the interruption, phase, and magnitude apprises the distribution of force on the surface. The main advantages of this approach are immunity of the optical sensors to high electromagnetic disturbances, providing a high spatial resolution, flexibility and durability with high speed of signal transmission.

In earlier work in [35], the optical tactile sensor is developed to measure the displacement and surface roughness with a high spatial resolution using artificial neural networks. One of the main approaches in optical sensing is to place optical fibres in the finger membrane and employ techniques like intensity modulation, Bragg grating, and specklegram. In [36], a Fiber Bragg Grating (FBG) sensor is proposed which has high sensitivity, but low spatial resolution of 5mm.

Another FBG-based tactile instrument is suggested in [37] to map the force distribution with a minimum weight sensitivity of 0.05 kg. A new class of optical tactile sensors are presented in [38] considering PolyDiMethylSiloxane (PDMS). A high sensitivity for measuring the minimum weight of 0.005 kg is demonstrated practically. Furthermore, a technique is offered to detect object shape and surface roughness with the sensors. In [39], a multi-modal optical tactile sensors combined with electrical tactile sensors demonstrate a successful application for material classification and proximity range detection.

A hybrid optical tactile sensor is proposed in [40] to measure normal force by employing Fluorescence. A camera with 41FPS is considered to capture intensity changes of the contact area proportional to the contact force. This sensor take advantage of the difference in wavelength excitation and Fluorescence emission which achieves a high signal-noise-ratio. The results indicate a high correlation of 0.986 with 3% error in reputability.

Most of the optical tactile sensors have a lower spatial resolution compare to vision-based techniques using cameras that are discussed in the next sub-section. Moreover, optical sensors require wiring through the silicone membrane which requires much effort to be adapted for different applications. Table 2.1 summarises different techniques for tactile sensing applications.

Table 2.1: A list of different tactile sensing techniques for various applications

Reference	Sensor	Purpose	Specifications
[20]	Capacitive	Measure the the force vector (3D)	• Low parasitic capacitance effect • Resolution of 12.5 μm • A considerable hysteresis
[31]	Capacitive and actuators	Object classification	• High accuracy of object recognition • Non-time series machine learning technique
[28]	Piezoelectric transducers	Classify material	• Time and frequency domain analysis • High accuracy with static applied force threshold
[25]	Piezoresistive	Measure the pressure	• High linearity factor • 100 cycles hysteresis • Time analysis of the measurements
[24]	Laser and piezoresistive	Object recognition and orientation detection	• High classification accuracy • Limited orientation measurements • Non-time series machine learning technique
[35]	Optical fiber	Measure displacement and surface roughness	• Measuring range of $\pm(0.8\ mm)$ • Displacement error of $\pm(0.5\ \mu m)$
[38]	Optical fiber	Measure the force and shape detection	• High sensitivity of 0.005 kg • Latency of 600 ms • Limited range of few grams

2.1.2.1 Vision-Based Tactile Sensors

Vision-based sensors are a subcategory of optical sensors that utilise cameras and image processing techniques for measurement and sensing purposes. Cameras are convenient instruments which are widely available in the market at a low-cost. Cameras have high resolution which enables vision-based sensors to be robust for precision applications [41]. They are easy to de-

ploy and integrate with external systems using widely available ports such as USB/LAN. Also, cameras can be replaced easily to change the sensor's resolution, speed and size in regard to the application.

Since image processing and machine learning techniques have been significantly advanced, vision-based approaches have become more popular in many applications. Moreover, embedded devices have higher processing and memory capabilities nowadays which can run multiple algorithms in real-time. All these factors add together to utilize cameras as a sensory instrument, known as Vision-Based Measurement (VBM) instruments [42]. The main principle of VBM instruments is to capture an image and perform algorithms on the image to measure or detect a physical parameter. VBM instruments have shown a remarkable success in various areas such as surgical tool detection [43], sign-based human machine interaction [44], roughness estimation [45] and navigation systems [46].

In tactile sensing, the fundamental approach of VBM sensors is to monitor changes within the contact area between an object and a soft membrane. Physical phenomena such as normal force and shear force on the surface results in the deformation of the soft membrane. Consequently, the visual features of the soft membrane change significantly which can be captured by the camera. These visual features can be processed to detect object slippage, identify texture, and estimate the contact force. Figure 2.2 presents a typical vision-based tactile sensor including a soft membrane and a camera.

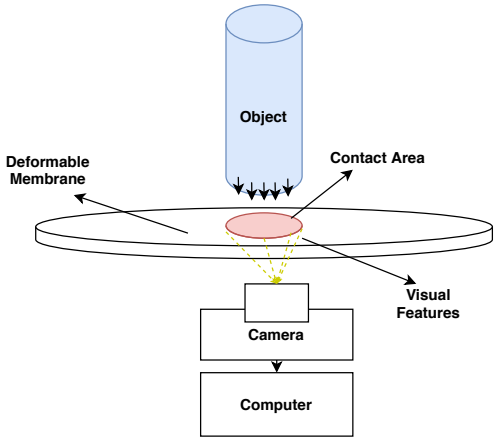


Figure 2.2: Diagram of a typical vision-based tactile sensor inspired by [47]

Vision-based tactile sensors are divided into three main categories based on the soft membrane design: (i) Reflective membrane; (ii) Light conductive plate; (iii) Marker-based membrane.

Reflective membrane sensors consider a reflective flexible material to observe changes in the shape of the membrane which enable the sensor to measure orientation as well as the texture of the object. One of the well-known vision-based tactile sensors in the field is GelSight that

employ a reflective flexible membrane, a conventional camera and lights [48]. The reflective membrane is built from Thermoplastic Elastomer with shore hardness between 5N to 20N. The high flexibility of the membrane, coupled with multi-colour lightening, allows the sensor to extract a high resolution of contact area shape from the images.

Light-conductive tactile sensors emit light to the contact area between the object and membrane. The camera is employed to acquire tactile information from the intensity changes. Popular materials for the membrane are acryle, glass, and silicone rubber which are widely available at a low cost. Light-conductive sensors observe the contact area shape either directly or indirectly. In the direct approach, the membrane reflects all the lights until the object touches the membrane. When the pressure is applied, the reflection of light reduces and the object's contact area becomes visible to the sensor. Indirect approaches follow the same principles while considering additional elastic layer between the object and the membrane [49]. In fact, the main advantage of light-conductive methods is that both hard and elastic materials can be employed for direct and indirect sensors respectively. In one of the earliest work [50], a vision-based tactile sensor is proposed which consists of a transparent acrylic plate, elastic sheets, halogen lamps and photo-transistor array. This study shows that the contact force has a relation with illuminance that are back-scattered to the photo-transistors. In [51], optical fibres have covered the surface of finger shape membrane which allow the camera to capture light illuminance changes when physical contact occurs.

Marker-based approaches embed small markers within the soft membrane which are displaced by the physical contact [52]. In this approach, the design of the soft membrane as well as the orientation of markers affect the sensor accuracy directly. The spatial resolution of such a sensor depends on the size and number of markers that are placed within the membrane. The most common shapes of the soft membrane are flat and hemisphere. The flat membranes increase the space between the grippers which allow the system to grasp larger objects. On the other hand, hemispherical soft membrane enables the sensor to be adapted for fingertips with curved surfaces [53].

Other approaches combine the aforementioned principles to acquire tactile information. For example, a light conductive plane and infrared LED are used in the sensor [54, 55]. This sensor consists of a stereo-camera to capture visible-light range and one camera for infrared lights. The stereo-camera is employed to measure the proximity while infrared lights provide a high spatial resolution for the contact shape.

In [56], marker-based and reflective membranes are combined to localize the object as well as the contact force measurements. The reflective membrane provides a clear image for the object

localization while the force measurements are estimated from the displacement of the markers. The sensor has a range of 0-10N and 0-2N for the normal and shear force respectively. Authors show the application of such a sensor in robotic bolt insertion and tightening. However, the image processing algorithms are based on object shape and they need to be tuned for each object. Moreover, although the camera has a sampling rate of 60FPS, the processing time of the localisation algorithm is reported as 3 seconds.

Similarly, a multi-modal approach based on reflective and marked membranes is proposed in [57] that measures object slippage in a robotic grasp. A transparent elastomer with markers is mounted under the GelSight reflective membrane. The markers are tracked during a grasp to measure the shear force and partial slippage. However, a tracking algorithm is designed for structured experiments and development of a general tracking algorithm for a grasp with different speed and angles will be challenging. Furthermore, the camera sampling rate affects the tracking performance for high-speed robotic grasping systems.

An interesting approach is proposed in [58] which considers a thermochromic liquid crystal ink layer with grated silicone rubber to measure temperature as well as the contact force. The contact force is calculated based on the displacement of grates on the rubber surface using Fourier Transform Profilometry. Also, the thermochromic liquid crystal layer reacts to different temperatures due to the change of the distance between the liquid molecules. This change of the distance results in the change of back-scattered light wavelength, and therefore, different colours are reflected for varying temperature on the surface. The temperature sensor has a limited range between 25°C and 31°C and infers the contact force calculations slightly.

All the vision-based algorithms aim to acquire tactile information based on the observed changes in the scene. For instance, light-conductive methods monitor the changes in back-scattered light while the marker-based methods monitor the displacement of the markers (changes in position of markers). Therefore, image processing techniques such as tracking and optical flow are developed to measure the motion within the contact area in [59]. However, these algorithms require a camera with high sampling rate to perform accurately for high-speed movements or changes in the scene. In addition to speed, the dynamic range of conventional cameras limits the sensor to capture changes within the contact area. Although high-speed and advanced cameras are available in the market, they are expensive and large which are not suitable for tactile sensing applications.

2.2 Neuromorphic Vision Sensors

In this thesis, a novel approach is proposed based on neuromorphic vision sensors and machine learning to tackle the limitations of the vision-based sensors. This section provides an overview of neuromorphic vision sensors and event-based applications. Figure 2.3 presents the common methods in event-based applications.

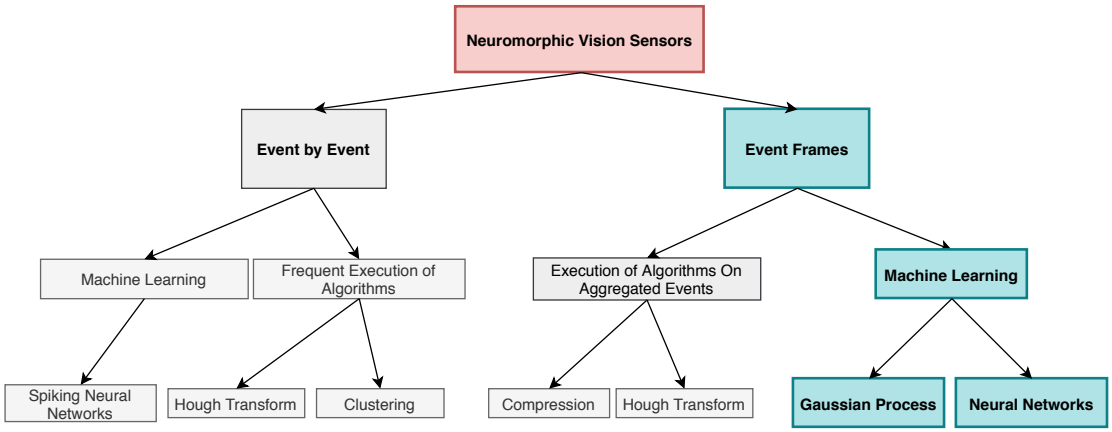


Figure 2.3: Diagram of neuromorphic vision-based approaches for different applications where the methodologies used in this thesis are highlighted in blue color.

Neuromorphic vision sensors (event cameras) are bio-inspired cameras that mimic the human's eye neural function. In opposed to conventional cameras, neuromorphic vision sensors capture the dynamic (intensity) changes of the scene asynchronously rather than the absolute intensity values synchronously. The asynchronous architecture of neuromorphic vision sensors and high bandwidth enables the sensor to offer much higher temporal resolution compared to conventional cameras. Figure 2.4 demonstrates a simplified circuit diagram of each pixel in DVS cameras. At first, a photo-receptor captures the light intensity which is scaled logarithmically. Then, a differentiator takes the difference between the current and the previous value. Afterwards, the comparator components compare the difference with a fixed threshold value. Finally, events are triggered if the difference exceeds the threshold, for positive and negative polarities.

The event camera streams information of each triggered event which includes location (x,y), timestamp and polarity (ON,OFF). The main advantages of the event cameras are the high temporal resolution (microseconds), high dynamic range and low power consumption. Therefore, these cameras are utilised for high-speed applications in challenging environments where the conventional cameras are not appropriate. Since the output of these cameras, i.e. stream of events, are fundamentally different from the image output of conventional cameras, the standard image-based applications are required to be redesigned for the event-based cameras.

Event-based applications are suitable for the dynamic environments as the sensor captures the

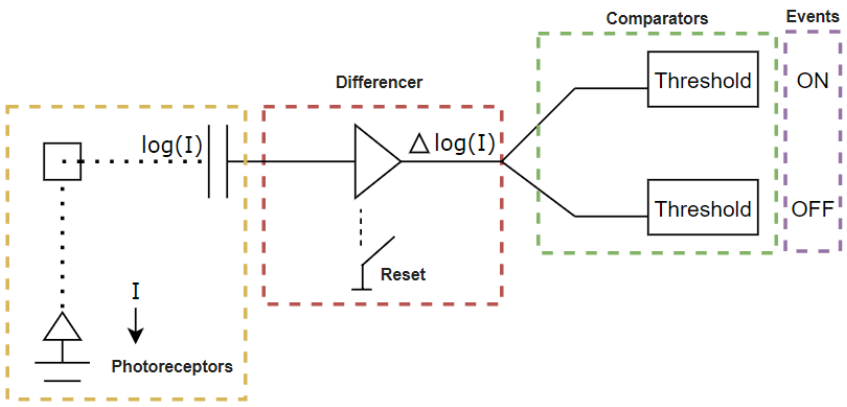


Figure 2.4: Simiplified circuit diagram for each pixel in DVS cameras, figure reproduced from [60]

intensity changes over time. As an example, the sensor mounted on autonomous cars or robots to capture changes in the scene while the system is moving. The movement of the sensor results in the changes of intensity in most of the pixels. Therefore, a significant number of pixels trigger events which can be used to identify textures and acquire further information from the scene. As events represents the intensity changes, the system is required to be designed with a memory to correlate events over time. This concept is similar to the video processing and motion analysis where both temporal and spatial features are necessary to design an algorithm.

Event-based applications consider two main approaches to process events [61]: (i) event-by-event; (ii) event groups (event-frame). The former approach processes events individually over time, while the latter integrates events over a time window to create a frame. Both approaches have been considered in recent event-based applications which are reviewed in the following section.

2.2.1 Event-based Camera Applications

Event cameras offer a higher temporal resolution of few microseconds compared to the conventional cameras. Hence, most of the event-based applications are designed to tackle vision-based challenges for high-speed systems. Moreover, the high dynamic range of event cameras makes them capable of capturing events in an environment with low light conditions. In this section, the main applications of DVS in robotics for the purpose of detection, tracking, and control systems are provided.

A robotic goalie system is proposed in [62, 63] with a low reaction time of 3ms for the known object size using cluster tracker algorithm. The clustering algorithm runs repeatedly on the triggered events which reduce the memory and processing requirements of the system. Afterwards,

a tracker is developed to track the clusters generated from the ball movement to feedback the controllers in real-time. The same system with a conventional camera requires at least 500Hz sampling rate as well as more memory and processing power.

In [64, 65], two event-based cameras (DVS) are employed to balance the pencil vertically from the triggered events. Events are processed individually to estimate pencil line parameters in the Hough space. Then, the line slope is estimated to feedback the controllers for balancing. The intervals of processing events are below than 1ms which is significantly lower than traditional frame rates. Thus, the system has sufficient time to detect slope changes and control pencil balancing.

Computer vision algorithms have been widely applied to high-speed navigation systems for Unmanned Aerials Vehicles (UAVs) [66] and autonomous cars [67]. Speed, power consumption and cost are the main important factors to design a high-speed navigation system. Therefore, a lot of studies consider event-based vision sensors as an instrument for measurements and analysis of the scene. For example, analysis of flight safety in unknown environments in regard to safety is reported in [68]. The study shows that event-cameras have a better advantage for the dynamic scenes compared to mono and stereo cameras. In [69], events spikes are processed to design a neuromorphic control system with Spiking Neural Networks (SNN) to control high-speed landings. The designed system process events one-by-one in all the stages.

On the other hand, several studies consider a group of events to extract features and build applications. In [70], the triggered events within a 3ms time-window are used to control quadrotors. The system has a low latency of 12ms which can track the line on the surface by using Hough transform and a Kalman filter. The results indicate that event cameras can be used in a closed-loop controller to achieve a system with a low response time.

In [71], an event-based system is proposed to estimate steering for self-driving cars. The events are integrated over a time window of 50ms to create frames (event-frames). Afterwards, event-frames are used to train a deep neural network to estimate the steering. The results indicate that the network trained on event frames outperforms networks using conventional images.

One may question that conventional cameras can capture frames every 50ms too, and therefore, subtraction of consecutive frames would result in the same output as the event-frames of the previous application. In such a case, the advantage of neuromorphic vision sensors is the higher dynamic range (140dB) compared to conventional cameras (60dB). As a consequence, event-frames can capture small intensity changes that conventional cameras may not, or operate in scenarios that conventional cameras suffer saturation. Furthermore, aggregated events present

the complete history of intensity changes over the threshold value within a time-window, while such information may be lost in conventional cameras, as each frame represents the the average intensity value during the capture time.

The events can have either positive or negative polarity based on the changes in the intensity. In each pixel, the higher intensity values in comparison to previous timestamp triggers positive polarity events. In contrast, the lower level of intensity level compared to the previous timestamp fires negative polarity events. Since the light conditions in the environment affect the polarity of the events, the polarity information can be ignored to generalise the algorithm for various environments. However, controlling the experimental setup and light conditions assist the system to acquire further information from the polarity values.

In this thesis, the experimental setup is designed to identify the changes in the contact force based on the polarity of events. The main elements to control this process are colour of the objects and camera position in the experiments. The positive polarity events and negative polarity events represent the increase and decrease of the object distance from the camera respectively. Therefore, the events are considered as two independent variables (two channels) for the development of the models.

As the event-based cameras have been widely used in mobile and robotic platforms, the efficiency of transferring and processing events plays an important role in the system performance. In [72], the events are downsampled spatially and temporally for the visual classification task. At first, the events are accumulated over a time window to create event-frames. Both spatial and temporal downsampling is performed on a single channel. A classifier is developed to investigate the effect of the time window and frame size on the performance for different applications like handwritten digit recognition and object classification. The results indicate that there is a trade-off between the time window size and performance of the classifier. Interestingly, a time window of 10ms increases the classifier accuracy significantly compared to the smaller time windows for the Caltech101 dataset considering low spatial downsampling.

In addition, the accumulation of events over time is used for the compression of the stream. In [73, 74], a compression technique based on the time aggregation of the events is proposed to reduce the data rate. Authors demonstrate that the compression ratio increases for a wider time window aggregation. In fact, increasing the time window aggregation up to a threshold improves both classifiers accuracy and compression ratio. However, there is a trade-off between the time window size and latency of the system in real-time applications like sensors which should be taken into the account.

As concluded in Section 2.1.2.1, vision-based tactile sensors are mainly limited by dynamic range and speed. On the other hand, event-based applications consider neuromorphic cameras based on four main factors: (i) Speed; (ii) Sensitivity; (iii) Memory requirements; (iv) Power consumption. The studies in the literature indicate that event cameras offer a high temporal resolution with low power consumption. The high dynamic range of event cameras provides a high sensitivity even in challenging light conditions. Additionally, accumulation of events over time reduces the memory requirements of the system. Therefore, event cameras are a suitable candidate as an instrument to tackle challenges in vision-based tactile sensors. Since the event cameras only capture the dynamics of the scene, image processing algorithms that are developed for conventional cameras cannot be applied directly on the event-frames. The relationship between the intensity changes and membrane deformation is very complex which cannot be modelled by traditional computer vision methods. In Section 2.3, relevant machine learning methods are reviewed to identify suitable modelling technique for this problem.

2.3 Machine Learning

In this Section, an overview of machine learning methods is presented. Figure 2.5 presents the structure of the review while the relevant methods to this thesis are highlighted in blue.

Machine learning is a data-driven approach to solve a problem by learning from experience [75]. The first step of designing a machine learning model is to define the problem as a function which maps the inputs and outputs of the system while the system accuracy can be measured by another function (loss function). For example, a vision-based tactile sensor receives a sequence of images (inputs) and provides tactile information such as force measurements. The loss function for this system can be considered as an error between the estimated force and the true force measurements.

Machine learning algorithms aim to optimize the loss function by observing examples (training data) and tuning the parameters based on the loss function iteratively. In general, increasing the number of examples results in the improvement of the machine learning model. However, other factors such as balance, diversity, groundtruth quality and data split ratio for evaluation of models affect the model performance significantly [76, 77].

Machine learning methods can be divided into three main categories: (i) Supervised learning; (ii) Unsupervised learning; (iii) Reinforcement learning. Supervised learning tasks require desired outputs (labelled) for each set of inputs while unsupervised learning intends to deal with unlabeled outputs (i.e clustering tasks). Reinforcement learning considers a different approach

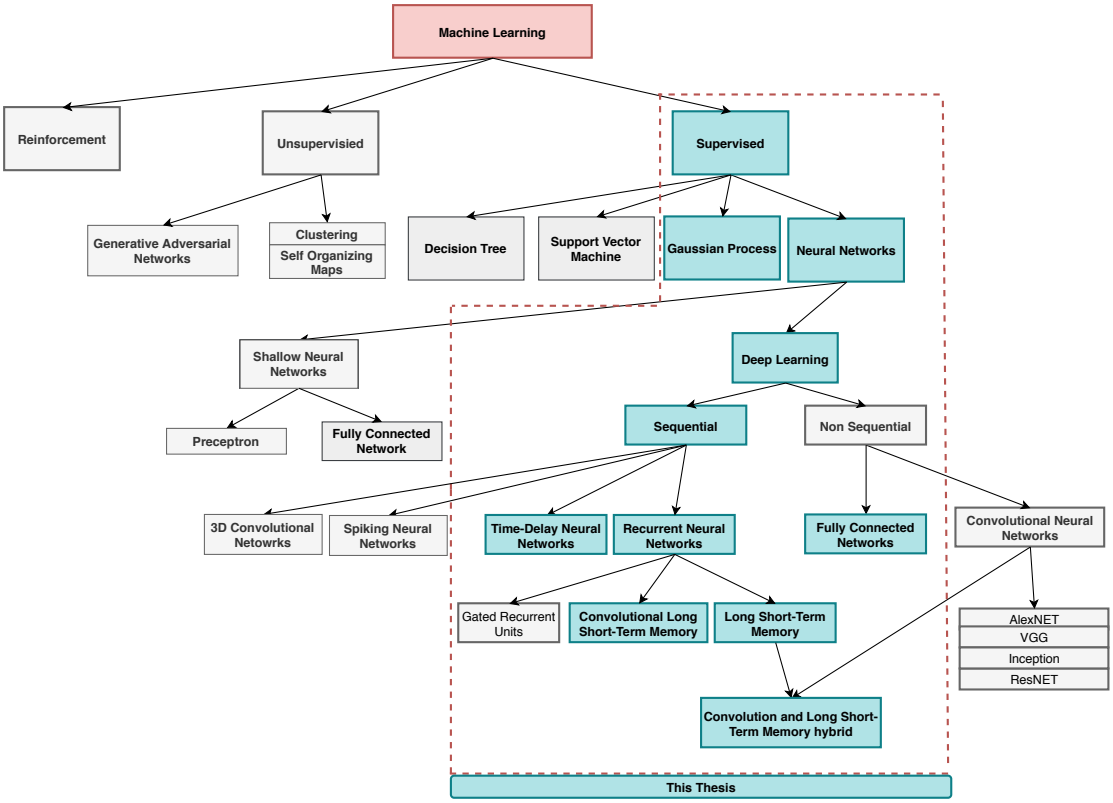


Figure 2.5: Diagram of machine learning methods. Blue blocks present related methods in this thesis.

to receive feedback from the environment to define a reward for the learning algorithm to choose the best possible actions sequentially. This research concerns the supervised learning methods to develop a novel vision-based tactile sensor for grasping applications.

In addition, machine learning applications are divided into two categories: (i) Classification; (ii) Regression. In a classification task, the machine learning model is designed to predict a label corresponds to a give input. For example, an image or video is provided as an input to the model and the model predicts the material of the existing object in the image. In a regression task [78], the model is designed to predict desired values based on a given input. For example, an image contact area between a soft object and a hard object is given and the contact force is predicted.

Moreover, the problems can be approached as sequential (time-series/videos) and non-sequential (images) tasks. Sequential learning adapts feature extraction as a function of time such as Recurrent Neural Networks (RNN) and Time-Delay Neural Networks. Therefore, the network learns temporal relationships between the inputs and the outputs by receiving a feedback from previous or future time steps.

Among the machine learning methods, deep neural networks and Gaussian processes have been

shown remarkable achievements in computer vision. In the following Sections, different types of neural networks and Gaussian process techniques are presented.

2.3.1 Shallow Neural Networks

Artificial Neural Networks (ANN) are bio-inspired machine learning models that intend to mimic animals and human brain structures. In early work, a learning algorithm is proposed to solve binary classification tasks. This algorithm is known as "Perceptron" which considers multiple inputs with a single output to classify two different categories [79].

Several other techniques are developed to adapt and improve the Perceptron algorithm for different applications by enhancing learning algorithms, increasing the number of nodes and using various activation functions. One of the well-known types of neural networks is "Shallow" neural networks where multiple neurons are stacked in a small number of hidden layers to regress or classify multiple inputs with multiple outputs. In addition to the number of neurons, the connections between neurons impact the network accuracy. Fully-connected network is the most popular network which has a connection for all possible pairs of neurons between consecutive layers. Such connectivity architecture allows each neuron to learn from all the input values during the training. For instance, in [80] features of medical images are handcrafted and passed to a shallow fully-connected network to segment regions of interest. Shallow networks have been used widely in the past for both regression and classification problems.

The main advantages of shallow networks are fast training procedures, a low number of hyper-parameters and capability of capturing simple relationships. Due to the limited number of hyper-parameters and neurons, shallow networks are unable to achieve accurate results when inputs and outputs of a network have a highly non-linear relation.

2.3.2 Deep Neural Networks

The recent development of hardware and learning algorithms provided this opportunity to implement networks with a significant number of neurons and layers [81]. Deep Neural Networks (DNN), also referred to as "Deep Learning", are multi-layer networks that have enormous number of trainable neurons. Due to the significant number of hyper-parameters, DNNs are capable of learning a wider range of features and relations compared to the shallow networks. Over last decade, enormous number of deep learning architectures have been designed to address a diverse range of problems in the computer vision. Auto-Encoders [82], Deep Belief Nets [83], Deep

Boltzman Machines [84], Generative Adversarial Networks (GAN) [85] have achieved remarkable results which are reviewed in [86]. Due to the dynamic nature of neuromorphic vision sensors, this thesis mainly considers relevant architectures for sequential and hybrid neural networks.

In the following sections, evolutionary networks in computer vision field such as convolutional neural networks, recurrent neural networks and time-delay neural networks are reviewed. The most impacting research are also highlighted.

2.3.2.1 Convolutional Neural Networks

Convolution is a mathematical operation that is mainly used in signal processing to implement filters by defining convolution matrices (kernels). Traditional image processing techniques rely on convolution operations for most of the feature extraction and filtering methods such as edge detectors, morphological operations, sharpen filters, etc.

Combination of convolution operations and neural network made a significant breakthrough in computer vision. Convolutional Neural Network (CNN) are bio-inspired non-sequential models that stimulate visual cortex of animals where each cortical neuron processes respective fields. Convolutional layers extract regional features in the image based on defined kernels. Before revolutionary CNNs, kernels and filters were designed by specialists for each task to extract useful features from images. However, this approach is changed by considering several convolutional layers in a deep network that are trained using appropriate data.

A typical CNN includes convolutional layers, pooling layers and fully-connected layers each of which is designed to perform a specific task [87]. Convolutional layers extract features by applying multiple filters in different sizes. Pooling layers are responsible for dimensionality reduction of the input by down-sampling techniques such as maximum and minimum of the neighbourhood pixels. Finally, fully-connected layers are considered to relate the extracted features to the desired output.

CNNs have become a golden standard among researchers after a significant success of AlexNet in ImageNet classification challenge 2012 [88]. AlexNet significantly improved the ImageNet classification accuracy with the top five error rate of 17% to classify 1000 categories. This network includes five convolutional layers, max pooling, drop out and three fully-connect layers.

Later on, a new architecture considering more layers and smaller convolutional kernels were proposed [89], referred to *VGG*. Stack of many small size convolutional layers with non-linear activation functions improves the discrimination of features through the network. This ma-

jor advantage led the network to perform robustly against most of the other architectures for classification problem in ImageNet Challenge 2014.

In contrast to the typical CNNs, a new type of layers (*Inception*) is proposed which considers the various size of convolutional kernels at each layer [90]. Multiple convolution with different kernel size at each layer enables the network to extract further features regarding the size of the kernels. Figure 2.6 presents the architecture of inception module. Multiple convolutions at

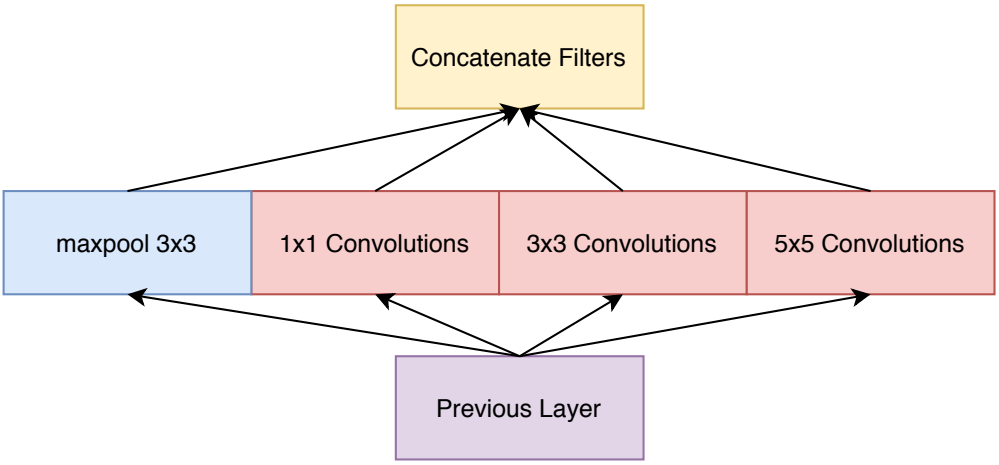


Figure 2.6: Inception module architecture regenerated from [90]

each layer may increase the computation cost of the deep networks significantly. Therefore, an alternative Inception module is introduced which reduces the dimension of data by adding 1x1 convolutional layers before main convolutions.

In addition to convolutional operations in layers, connections of the layers and learning algorithms are of great importance in a network. A new type of network, known as "ResNet", is suggested [91] which considers shortcut connections between convolutional layers (Residual Layers). The identity shortcuts allow the network to consider the main input as well as the processed images at each layer. Interestingly, it is demonstrated that this type of network can be trained over 1202 layers without any optimization problem.

To increase the sensitivity of the network to informative features, a new method *Squeeze and Excitation* is proposed with consideration of global average pooling for local descriptors at each channel. To capture dependencies over channels, a non-linear gating method is considered to correlate channels together.

CNN architectures have shown significant success in non-sequential applications [92]. For the sequential data like videos, the same principle of 2D convolutional layers applies in 3D, known as 3DCNNs [93]. Since the kernels are performed on both spatial and time domains, 3DCNNs are robust to capture short spatio-temporal dependencies. For long-term dependencies, the

CNNs are often combined with Recurrent Neural Network (RNN). In the following Section, an overview of the most impactful sequential networks including Time Delay Neural Networks and RNN architectures are provided.

2.3.2.2 Time Delay Neural Networks

Time Delay Neural Network (TDNN) is another type of networks for sequential data points. Similar to RNNs, TDNNs models are mainly designed to capture dependencies of the data points with past observations [94]. The main advantage of TDNNs is the ability to relate temporal sequences to each other, enabled by their main characteristic and the delay nodes. The number of delay nodes is a crucial parameter in the TDNN network which specifies a time interval to capture patterns of a signal.

In addition, TDDNs requires fewer computations and often have faster training time compared to RNNs due to the constant value of the delay nodes. The delayed node in TDDNs is often assigned to a short time to capture the relationship of the current point with the previous observations. A fixed and small number of time delay nodes helps to avoid vanishing gradient problem as mentioned in Section (2.3.2.3). Since TDDNs are suitable for short-term dependencies, the representation of inputs can be adapted with a memory to capture long-dependencies. In Chapter 3, the effectiveness of TDNNs parameters by adapting the inputs is investigated for a dynamic sensor.

2.3.2.3 Recurrent Neural Networks

Recurrent Neural Network (RNN) aims to solve sequential problems [81]. Unlike traditional neural networks, RNNs behave dynamically which allow learning model to have a finite memory. RNNs have been used widely for time-series regressions and classification problems such as speech recognition [95], action recognition [96], and time-series predictions [97].

The main principle of RNNs is to provide feedback after each timestamp which provides memory to the network. Traditional RNNs made remarkable progress to improve learning performance for sequential data. However, long-term dependencies of variables cause a significant problem, known as "Vanishing Gradient" problem [98]. The main reason for the vanishing gradient for traditional RNNs is that the feedback of each timestamp is not controlled. Therefore, the recent timestamps have a significant impact on the training process which causes the network to forget long-term dependencies. The vanishing gradient problem is also observed in very deep

networks where the data fades in very deeper layers, and residual layers were designed to address this problem which is discussed in Section 2.3.2.1.

To address the vanishing problems for RNNs, a new type of units is introduced for long and short-term dependencies. Long-Short Term Memory (LSTM) unit consists of a memory cell and three gates which assist the network to keep the meaningful dependencies and remove unnecessary features over sequences [99]. Figure 2.7 presents an LSTM module where X_t. C_t, h_t represents the input, memory cell and hidden state respectively, where the subscript t represents the timestamp.

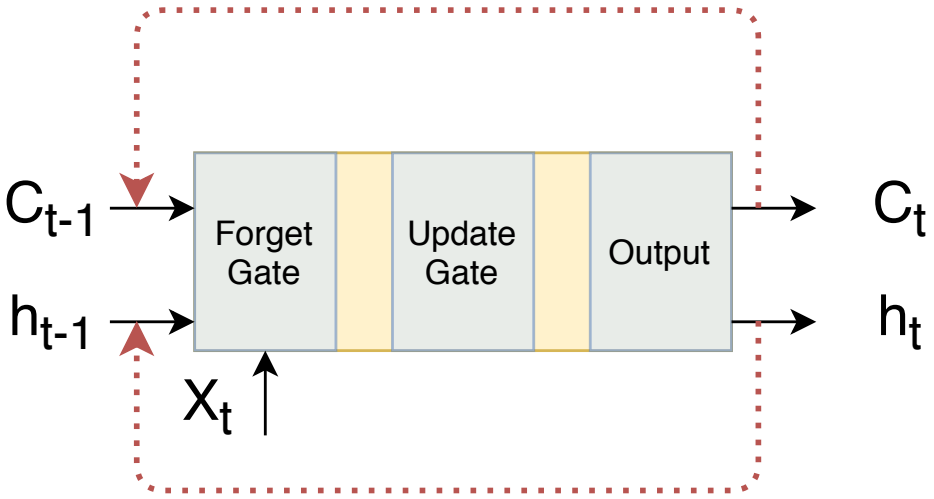


Figure 2.7: Representation of a LSTM module with forget gate.

In a dynamic sensor (e.g. neuromorphic vision sensor), only the intensity changes are measured in the scene at each pixel. In fact, the absolute intensity of the pixel is not directly measured and can only be approximated through the whole history of measured changes. Assuming a vision-based sensor, the tactile information depends on the deformation of the membrane in the contact area. Hence, the system must capture dependencies from the start of the experiment to evaluate the current state of the contact area.

For the sequential data, LSTMs are robust to capture both long-term and short-term dependencies which makes them an appropriate candidate for dynamic sensors. In a tactile sensor, long-term dependencies are crucial for estimation of the contact force since the current value depends on the history of the force from start of the experiments for the dynamic sensors. On the other hand, short-term dependencies must be captured to identify the variation of the contact force in the recent frames where the network is able to learn incipient slippage point as well as vibration patterns in the experiment.

On the other hand, CNNs offer an accurate feature extraction in 2D and capturing short-term

dependencies in 3D. Therefore, a hybrid approach can be developed to extract spatial features by CNN and capture both long-term and short-term dependencies with LSTM layers. In Chapter 4, LSTM and hybrid approaches are investigated and discussed. Both LSTM and CNN hybrid approaches are computationally expensive which increase the training and inference time. In the following section, Spiking Neural Networks are presented

2.3.2.4 Spiking Neural Networks

As mentioned in Section (2.3.1), neural networks are bio-inspired learning models to mimic animals and human brain. Spiking Neural Network (SNN) is the third generation of networks which mimics the human brain closer than other networks. This generation of networks considers a potential threshold for each neuron to be activated. In contrast to other types of neural networks, neurons of SNNs have two states of "ON" and "OFF" based on the threshold value. Therefore, the neurons transmit information at activation time rather than each propagation cycle. The activated neurons send a new signal to other neurons in the network which decays over time. The time between the activation of neurons and the frequency of activated neurons are used to encode information of a network.

In recent years, SNNs have been in the centre of attention, especially for neuromorphic computing. Many sensors such as Dynamic Vision Sensors and Dynamic Audio Sensors as well as processors like TrueNorth illustrated capability of neuromorphic devices in different applications. Neuromorphic devices have been proven low latency and power consumption compared to other types of sensors for similar applications [60, 100]. Deep SNNs and combination of CNN with SNNs have achieved a competitive performance on well-known image benchmark datasets [101]. Moreover, SNNs illustrate a great performance by using neuromorphic sensors streams (event-by-event) [102]. However, algorithms and hardware for conventional cameras have been developed over decades with a remarkable success for various applications in practice. A lot of frameworks and hardware choices are available which accelerate the development of the sensor. In addition, there is a lack of unified frameworks and benchmarks for computations and models in spiking neural network field [103]. Therefore, event framing technique is considered in this thesis to create artificial frames from the events which is compatible with ordinary neural networks and GPUs.

2.3.3 Gaussian Process

Gaussian Process (GP) is a stochastic modelling technique to predict and forecast variables based on the combination of random variables over the data points. Increasing the number of random variables in this process enables GP to be robust against highly non-linear functions. A variety of kernels can be considered to fit a function with random variables corresponding to the multivariate normal distribution [104]. The choice of the kernel functions and hyper-parameters have a significant impact on the model to estimate the function between inputs and outputs.

In computer vision, GP has been widely used for both classification and regression problems. For instance, an interesting approach for human tracking is proposed in [105], which considers human joints position as a time-series regression task. It is worthy to note that modality of the data is an important parameter to develop an appropriate machine learning method. In [106], a new framework is introduced to improve the performance of the time-series GP classification model for sparse and uncertain sampled data.

Recently, an increasing number of studies investigate deep architecture learning models. A deep GP model is proposed in [107], where GP models are connected together. Each layer of GP produces new inputs for the next layer and Bayesian optimization is considered to train the model. However, it has been studied that GP loses its efficiency in high dimensional space by assuming a stationary covariance for all points in the inputs [108].

One major problem of machine learning methods is that the accuracy of machine learning models relies on the training data significantly. Therefore, any limitations in the training data reflect on the sensor accuracy in real-world scenarios. In chapter 5, augmentation techniques are reviewed which aim to generate training data artificially to improve the accuracy of machine learning models.

2.4 Limitations and Direction

This chapter reviewed related work in regard to the tactile sensors, neuromorphic vision sensors and machine learning methods. In section 2.1.1, non-optical tactile sensors were studied and it was shown that non-optical techniques are suffering from low spatial resolution.

Different categories of the vision-based sensor were reviewed in Section 2.1.2.1 which indicates that vision-based sensors offer a high spatial resolution while they are resistant to electromagnetic fields. However, vision-based sensors have a limited sampling rate and dynamic range, typically

30FPS and 60dB respectively. Cameras with high sampling rate are available commercially but these cameras are large, costly and have a high power consumption which are not suitable for robotic applications. Therefore, the main limitations of vision-based tactile sensors are as follows:

1. Limited sampling rate
2. Low dynamic range
3. High power consumption

In section 2.2, neuromorphic vision sensors (event cameras) were presented and various applications in literature were highlighted. Event cameras provide a high temporal resolution (few microseconds) with low power consumption. Nevertheless, the transformation of intensity changes to the tactile information is very complex which requires advanced machine learning techniques with capability of sequential modeling in time-domain.

In section 2.3, machine learning methods were studied and relevant research is highlighted. Since the event cameras only capture the dynamics of the scene, sequential machine learning models were reviewed. In a dynamic sensor, current measurements are depending on the history of measurements. In fact, the sensor must have a memory to relate dependencies between different timestamps. Therefore, machine learning techniques such as DNN, TDNN, GP, LSTM networks, and hybrid CNNLSTM are further investigated in this thesis. Since machine learning approaches are data-driven methods, lack of training data reduces the accuracy of the models significantly. In chapter 5, augmentation techniques are presented to synthesise training data and improve the models' accuracy without performing real experiments.

Chapter 3

Temporal Neuromorphic Tactile Sensing

3.1 Introduction

The main limitations of vision-based sensors were identified as camera sampling rate, dynamic range and resolution for marker-based methods. In this chapter, a novel tactile sensor based on a neuromorphic camera is proposed to acquire tactile information based on intensity changes of the contact area in order to estimate the contact force and classify objects' materials. The contributions of this chapter are: (i) Design of a novel neuromorphic vision-based tactile sensor;(ii) Development of Gaussian Process and Time-Delay Neural Networks to estimate the normal contact force for objects with the same shape and size using spatial features (iii) Development of a Deep Neural Netoworks (DNN) to classify objects' material in a grasp by considering spatial features. The results are validated on real-world experiments where the objects with the same shape and size are grasped and released after a short time. To generalise the sensor for objects with a different size, a novel method and experimental setup is proposed in chapter 4.

This chapter is structured as follows. In section 3.2, related studies in vision-based tactile sensors are reviewed. Design of the novel neuromorphic vision-based sensor is presented in section 3.3. The sensor operation principle and development of TDNN and GP are presented in section 3.4. In section 3.5, a DNN model is designed to classify objects' materials considering various elasticity. In section 3.8, an overview of findings is presented and the chapter is concluded.

3.2 Related Work

A general overview of various methodologies and modalities for tactile sensors was provided in chapter 2. In this section, vision-based tactile sensors for the contact force estimation and material classification are reviewed with a focus on the marker-based elasotmer methods.

Recent developments in visual technologies made cameras available in smaller sizes, lower cost, and higher resolution. Furthermore, advancements of processors and computational devices enabled the cameras to be considered as a Vision-Based Measurement (VBM) instrument to measure physical properties, localisation of the objects, and to classify materials [109].

Vision-based tactile sensors observe the contact area, object, and elastomer membrane surface to detect slippage and estimate the applied force. Different categories of the vision-based sensors were reviewed in chapter 2. The most popular method for measuring the contact force is marker-based vision sensors. In earlier work in [110], a camera and a force sensor are utilised to estimate the deformation of an elastic object and marginal slip. A marker is placed in the center of membrane to measure radius of the contact area while the force sensor measure the contact force. In another approach [52, 111], a vision-based tactile sensor is proposed, known as GelForce, to measure contact force in 3 dimensions by utilising a 30FPS camera and markers in the membrane. Two layers of 24×24 markers with different depth and colours are considered to calculate the deformation based on B-Spline wavelet transform [112].

Similarly, a membrane with dotted pattern on surface is designed in [113] to estimate the contact force and stick ratio of the surface. The results are compared to a PVDF sensor which indicate that the vision-based sensor achieves less noisy measurements for measuring shear-force. An extended version of this work is presented in [114] to detect micro-slippage in advance by quantisation of maximum markers displacement. Later on, the authors combined the functionalities of the sensor into a touch-pad that acquire object orientation, slippage, 3-D force measurements and the contact state [115]. Unlike other tracking-based methods, the authors proposed re-identification technique based on the distance between neighborhood markers. However, the results show that the contact state in each region of the membrane includes false-positive detection even when the membrane is deformed considerably.

A different approach in [116], considers features of objects' surface to detect slippage. Authors proposed a novel technique by using the Speeded Up Robust Features (SURF) algorithm to extract and match the features. However, the proposed sensor cannot be used for transparent or reflective materials due to the disturbance of light reflection. Moreover, a camera, a textured membrane and a light diode are used in [117] to compute force magnitude and find directions for

several rigid and soft contacts. The sensor has a range between 0-7N with resolution of 0.05N while each frame takes 55ms to be captured by camera. Similar marker-based approaches are used in [118–121] to measure the contact force in 3 dimensions. The marker-based approaches aim to transform the membrane deformation into displacement vectors based on the membrane strain. Therefore, the sensor is capable of measuring force in three dimensions.

In addition, researchers have attempted to adapt machine learning techniques for vision-based measurements devices. For example, a Recurrent Neural Network (RNN) is trained to estimate contact force using a pair of cameras [122, 123]. The authors demonstrated a low Mean Absolute Error (MAE) of 0.0464N for the contact force estimation. However, the experimental setup is controlled and the contact position of the tool is provided to the network. Furthermore, stereo-vision allows the system to capture depth of the scene which is a less challenging task compared to single camera methods due to the estimation of depth from stereo-cameras. Other stereo vision-based force measurement methods can be found in [124–126] for a variety of applications in tactile sensing field.

In addition to the force measurements, classification of materials assists the system to perform tasks in an adaptive manner. For instance, fragile objects can be classified in order to prevent breaking of the objects by controlling the contact force. In [127], a multi-modal approach is considered to classify objects using both tactile and vision information. Authors present that two-stream Convolutional Neural Networks (CNN) transforms the inputs into a latent space for material classification. Although the multi-modal approach achieves a high accuracy of 90% for texture recognition, the system has a poor performance for cross-modalities between tactile and vision sensor.

In [128], a MicroElectroMechanical System (MEMS) tactile array is used to classify objects in a single grasp. The results indicate that random forest learning method achieves 90% accuracy for classifying 11 objects grasped in different orientations. Although the sensor can perform well for objects of different shapes, the classification accuracy for objects of similar shape is reduced significantly. Moreover, the sensor is sensitive to electromagnetic field interference and has a limited resolution.

A hybrid method is developed in [129] which utilises a camera and a tactile sensor to acquire tactile information dynamically. Authors demonstrated the challenges in dynamic vision sensing such as vibration and noise which requires advanced modelling techniques.

As reviewed in chapter 2, most of the vision-based tactile sensors focus on the force measurement under stable and static conditions, i.e. without dynamic variation of the applied forces.

However, in many applications including robotic grasping, applied forces may vary significantly and a fast response is required to properly handle the grasped object. Even-though many VBM instruments and hybrid techniques have been contributing significantly in the field of tactile sensing, no attention has been paid to employ neuromorphic vision sensors in this field. For the first time, in [130], a neuromorphic vision-based sensor is utilised to detect incipient slip using Dynamic Vision Sensor (DVS) which provides a low latency with low power consumption. It is demonstrated that the sensor can detect incipient slip in grasping applications with an average of 44.1ms using traditional image processing methods. Although this sensor has shown the potential use of neuromorphic vision sensors for tactile sensing, it is unable to provide force measurements during the grasp. This thesis proposes a novel technique to estimate the contact force in a grasp which adds a further functionality to cover a wider range of applications.

In this chapter, a novel approach is proposed to estimate the contact force by using a neuromorphic vision sensor. Two machine learning techniques are implemented to learn the contact force measures from the triggered events. Furthermore, it is demonstrated that the objects with different Young's modulus can be classified after a grasp which allows the system to identify the objects' materials.

3.3 Neuromorphic Vision-Based Tactile Sensor

In chapter 2, neuromorphic vision sensors and their applications were presented. One of the well known neuromorphic vision sensors is DVS (Dynamic Vision Sensor) which has a high temporal resolution of few microseconds [60, 131], significantly faster than ordinary cameras. DVS is the first neuromorphic camera (invented by iniVation AG) that was available to research community from early stages. Recently, other manufacturers such as Samsung, Prophesee and CelePixel and AIT Austrian Institute of Technology have been producing neuromorphic vision sensors which are available commercially. In this thesis, DVS is chosen based on availability in the market and the support provided by iniVation AG to research community.

This vision sensor captures intensity changes logarithmically at each pixel rather than capturing the whole scene in a fixed interval. At each pixel, the intensity values are compared against the latest intensity values to trigger events (Figure 2.4). If the intensity level is increased, an event with a positive polarity (1) is triggered. In contrast, if the intensity level is decreased, an event with a negative polarity (0) is fired. Each event is characterised by location (x,y), timestamp and polarity. In this thesis, DAVIS 240C is utilised due to the availability of the sensor for the research groups. DVS has a resolution of 240×180 pixels with a latency of 12

microseconds for the mean of 20 events. The sensor streams positive and negative events with a precise timestamp and pixel location (x, y). Moreover, DVS requires lower power (4-15mW) and less memory compared to conventional cameras.

The threshold level of the events is a crucial parameter for filtering noise and changing the sensitivity of the sensor. Several factors such as environment lighting, distance of the objects, colour of the object and camera lens affect the optimal threshold of the sensor. In general, there is a trade-off between the threshold level of events and the sensor sensitivity. A low threshold results in the high sensitivity of the sensor as well as increasing the noise level of the scene. In this thesis, the threshold level is tuned to reduce the background noise by trial and error. The grasping procedure is repeated and the background noise is observed visually to ensure the minimum noise with the certain threshold. Most of the noise events are triggered with a positive polarity, and therefore, the threshold for positive events is defined higher than the threshold for negative events. Since the experiments are performed in various environments, the tuning process is repeated for each experimental setup to find an appropriate threshold level. In future work, the sensor will be covered by a case with a fixed lighting condition to eliminate the effect of external factors such as environmental lighting conditions on the sensor. Therefore, the tuning process can be performed only one time to achieve the optimal results in different environments.

The relationship between the triggered events and intensity changes is logarithmic which is formulated in [60]. Equation 3.1 presents the correlation of temporal contrast (TCON) and photo-current (I). A threshold is considered for the temporal contrast to fire positive (higher intensity) or negative (lower intensity) which can be modified to change the sensor sensitivity and filter the noise.

$$TCON = \frac{d(ln(I(t)))}{dt} \tag{3.1}$$

To establish the neuromorphic vision-based tactile sensor, a semi-transparent silicone membrane (light-conductive) is located between the object and gripper. When a force is applied on the object, the silicone deforms due to the elasticity effect. Therefore, DVS captures the intensity changes within the contact area and triggers both positive and negative events, as presented in Figure 3.2. In this chapter, the silicone membrane is molded with dimension of 4.0×2.0×0.2cm to cover the contact area between the gripper plane and the object. The dimension of silicone is considered based on the gripper size which can be adapted for different grippers.

The silicone properties and depth of the membrane have a significant impact on the sensor sensitivity and the range of force estimation. In fact, after a certain amount of increase in force,

the silicone membrane reaches a saturation point where further increase of the applied force does not deform the membrane considerably. Hence, the sensor can estimate a limited range of force which depends on the silicone membrane properties which is a case for all the vision-based tactile sensors with a elastic membrane. Figure 3.1 illustrates a diagram of the sensor including transparent grippers, a semi-transparent silicone membrane, DVS, and the object.

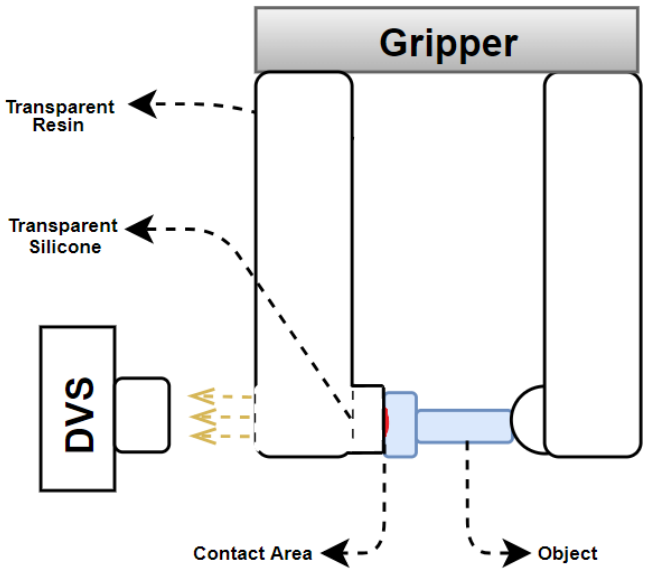


Figure 3.1: Event-based tactile sensor diagram

3.4 Force Estimation

In this section, two novel methods are proposed for the contact force estimation. In section 3.4.1, principles of the sensor operations are described. Afterwards, the grasping procedure is analysed in section 3.4.2. Then, Time-Delay Neural Networks (TDNN) and Gaussian Process (GP) methods are proposed in section 3.4.3 and section 3.4.4 respectively.

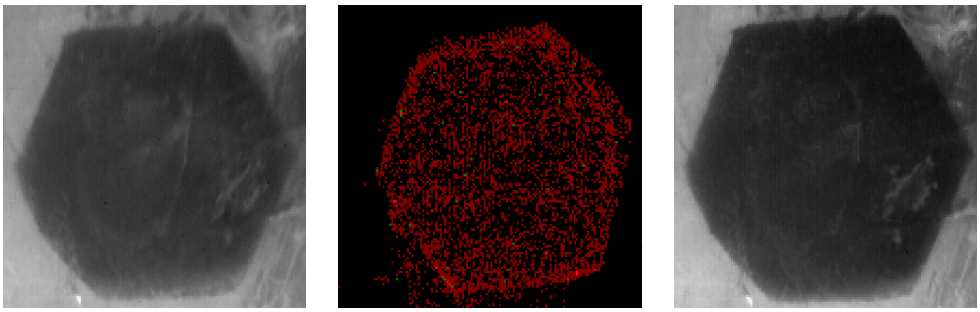
3.4.1 Sensor Operation Principles

As mentioned in section 3.3, DVS fires either positive or negative events depending on the intensity changes in the scene. Since the semi-transparent silicone membrane has an opaque surface, the contact area is barely visible prior to the contact of an object to the membrane. Due to the deformation of the silicone membrane, the visible part of the contact area becomes larger by applying more force, and intensity of the contact area increases significantly. Also, the intensity of the contact area is changed due to the reduction of distance between the contact

area and DVS. Accordingly, an increase of the applied force triggers the negative events while a decrease of the applied force triggers positive events. In this thesis, positive and negative events are presented in green and red respectively.

Deformation of silicone under a pressure is highly non-linear which depends on the type and size of the membrane as well as the range of the applied force. Other factors such as direction of force, shape of the contact area and temperature can affect this relationship. Consequently, the correlation between events and the contact force is highly non-linear considering the following parameters: (i) The deformation of silicone[132]; (ii) The logarithmic relation between changes in intensity and triggered events which is presented in Equation 3.1.

To visualise the correlation of triggered events to the contact force, events are accumulated over time interval where the applied force increases significantly. Figure 3.2 represents the triggered events and image of the contact area where the contact force is increased. Figure 3.2(a) and 3.2(c) are captured by active-pixel module of DVS to visualise the actual image of the contact area for demonstration purpose only. In Figure 3.2(b), the triggered events are accumulated over 40ms time window where the most regions of the contact area are covered by the events. On right bottom of the contact region, a number of events are triggered due to the noise and a slight displacement of the silicone membrane.



(a) (b) (c)

Figure 3.2: (a) Image of the contact area when a low amount of force is applied. (b) Accumulation of events over a 40ms time window during a grip. (c) Image of the contact area when a high amount of force is applied.

As observed in Figure 3.2, the proposed sensor captures the membrane deformation without markers in the membrane. In fact, each pixel plays the marker role in the membrane by triggering events to capture the deformation of membrane. In this chapter, the triggered events are accumulated from the start of the experiment by ignoring spatial information. Therefore, each frame just contains an integer number which presents the number of events that are triggered up to the current timestamp.

3.4.2 Grasping Procedure

The contact force estimation from DVS events can be approached as a time series regression problem. A single grasp can be divided into three main phases: (i) Grasping phase; (ii) Holding phase; (iii) Releasing phase. The contact force changes significantly in the grasping and releasing phases while in the holding phase the force variation is related to the vibration.

At the first instance when the object touches the membrane, a lot of negative events are triggered due to the intensity changes in all of the correspondent pixels of the contact area. Therefore, a first touch is determined when the first significant number of negative events are triggered. Once the contact is obtained, negative and positive events represent the changes in the applied force and vibration of the object.

In the holding phase, the applied force varies due to vibration and noise which requires a very high sensitivity to be detected. In this chapter, only the grasping and releasing phases are considered where a significant number of events are triggered. Finally, in the releasing phase, the applied force is decreased which leads to trigger positive events within the contact area. Since the threshold level of positive and negative events are varied, the accumulation of events for each polarity has a different peak. Therefore, the number of events and the contact force are normalised (maximum normalisation) to demonstrate the events and the contact force in Figure 3.3.

Both number of events and the measured force are framed over time intervals. The framing process helps to differentiate meaningful events and reduce the impact of noise over a longer period. In this chapter, only grasping and releasing phases are taken into account since the range of contact force is limited in the holding phase. In an ideal grip, the grasping phase must include only negative events. However, the object vibrates slightly in a short amount of time to reach stability which causes triggering the positive events. There is a trade-off between the filtering of the unwanted events and the sensor sensitivity which can be adjusted by changing the DVS threshold.

In Figure 3.3, the first peak in the negative events represents the first touch of the object and the membrane. Since applying force to the object reduces the distance between the object and camera, the intensity values of the pixels decrease dramatically. Therefore, a significant spike is occurred in negative events during the grasping phase. In contrast, the intensity level of the scene increases significantly during the releasing phase. Consequently, peaks of positive events are expected during the releasing phase. The threshold of the positive events is lower than the one of negative events which makes the sensor to have a higher sensitivity for the decrease of the

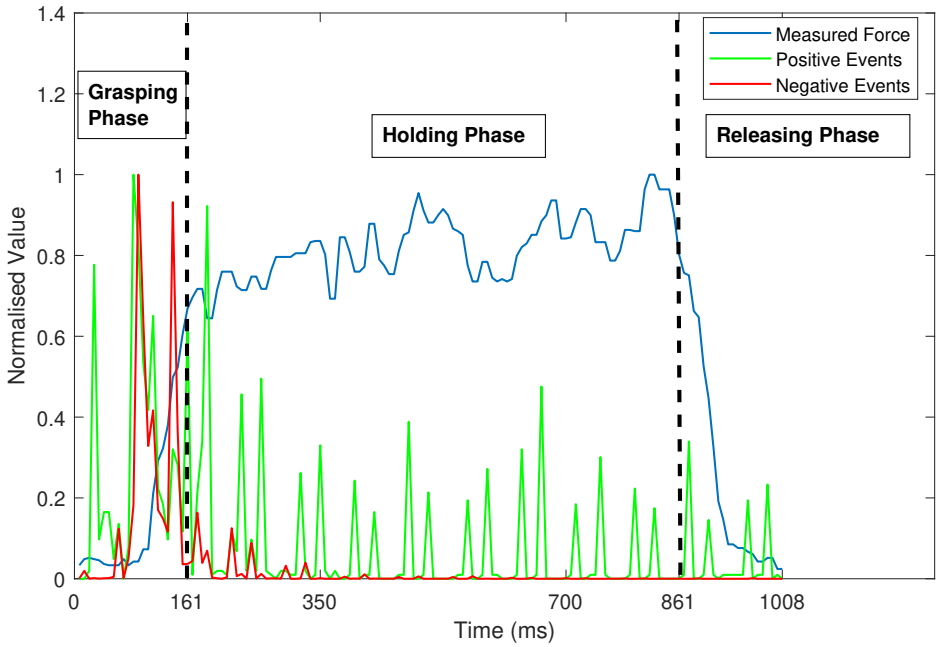


Figure 3.3: Normalised value of the applied force which is measured by a piezoresistive force sensor (blue), normalised number of negative events (red) and normalised number of positive events (green) in a single grasp.

contact force. In an ideal grasp, only negative events are expected to be triggered by applying more amount of force. However, vibration and instability of the grasp result in triggering positive events as well as negative events. The first significant spike in the releasing phase demonstrates a loss in the contact area which leads to the object slippage. Noticeably, positive events have a higher amplitude during the holding phase. The main reason of this phenomenon is that micro-vibrations change the intensity of the contact area significantly while the force sensor with a limited resolution cannot capture these vibrations. Moreover, the force readings are filtered to eliminate the noise which affect the force sensor sensitivity.

To correlate the triggered events and the contact force, a robust time-series learning technique is required to capture the non-linear relationship over a time. In this chapter, TDNN and GP models are implemented to estimate force during the grasping phase and the releasing phase. Both TDNN and GP models are designed to correlate inputs and outputs temporally which make these methods suitable for the contact force estimation. Accordingly, the force values are measured by a piezoresistive force sensor at each time interval to train and test the machine learning methods. It should be noted that the measured force values are used to train the models and the accumulation of triggered events are considered as the inputs to the models. In section 3.6, more details of the experimental setup are provided.

3.4.3 Time Delay Neural Networks

As mentioned in Section 2.3, TDNN are powerful learning models for sequential data. In this chapter, a variety of networks with different number of hidden layers and neurons are tested to find the best architecture. As discussed in chapter 2, TDNN capture the short-term features due to the delay nodes. Since the number of events are accumulated from the start of experiment, the history of triggered events are presented in the current timestamp. Therefore, long-term information is presented at each timestamp which will be captured by TDNN. It should be noted that this method would result in vanishing gradient for very long sequences as the accumulation of events will be significant.

The accumulation of positive and negative events are passed to the network separately to identify decrease and increase of the contact force respectively. Since polarity of the events indicates the direction of normal force variations, accumulation of positive and negative events are considered as independent inputs (features) of the network. In this chapter, the input has a dimension of 2×30 for two features of thirty timestamps. Furthermore, the input layer consists of four nodes for each variable including one node for the current and three nodes for the previous timestamps. Followed by the input layer, k fully-connected hidden layers with n neurons in each layer are considered to capture the non-linear relationship between the events and the applied force.

The sigmoid activation function is assigned to all hidden layers after some initial experimentation. A variety of experiments are performed to find appropriate parameters for the model to achieve a good performance. In section 3.7, a variety of network architectures are analysed comprehensively to investigate the impact of the number of neurons and hidden layers on the network performance. Figure 3.4 demonstrates the deep TDNN network for the force estimation.

Assuming the relationship between the input $x(t)$ and output $y(n)$ are mapped by the function H in Equation 3.2. TDNN aims to define H by optimising the weights (W_{ij}) of ith neuron in input layer to the jth neuron of the hidden layer as well as biases (b_j) of jth neuron in the hidden layer where the number of delay nodes are denoted as l in Equation 3.3. The function of input and output layers are denoted as net_j and y_n respectively. Finally, the output of the network is defined as a function of weight matrix (W_{ij}) and the activation function $\delta(x) = \frac{1}{1+e^{-x}}$ of input in Equation 3.4. The equations are adapted from [133] based on the sigmoid activation function and 3 delay nodes ($l = 3$).

$$y(n) = x(t+1) = H[x(t), x(t-1), x(t-2), x(t-3)] \tag{3.2}$$

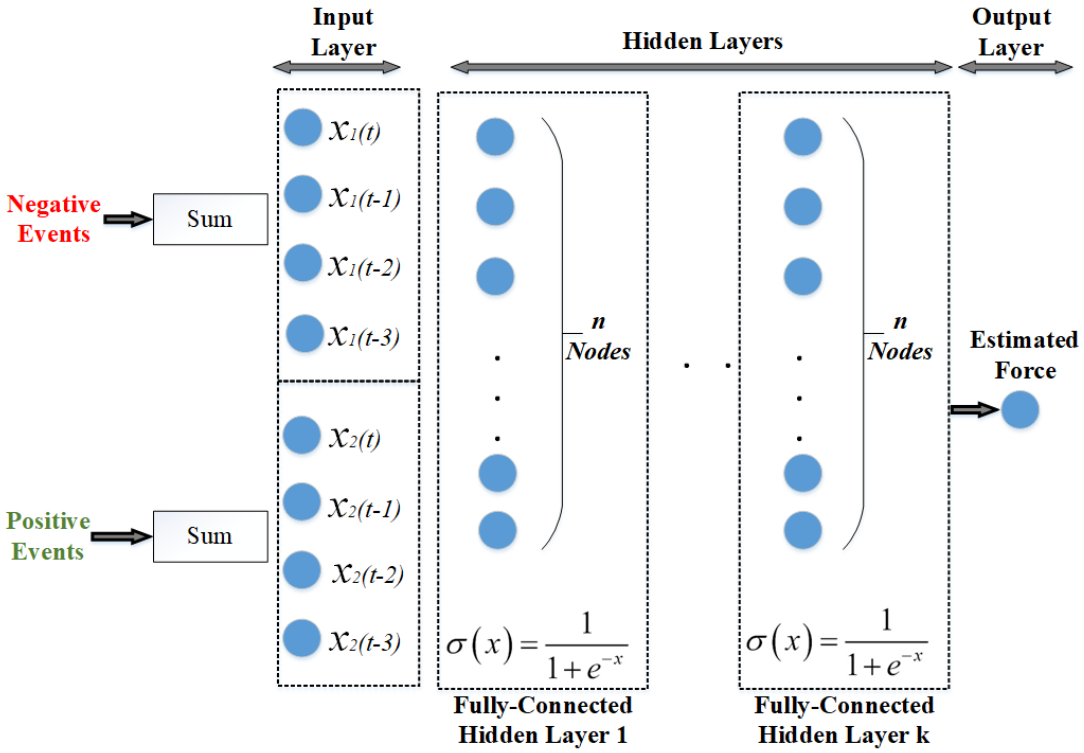


Figure 3.4: A deep TDNN model with a time delay of three nodes (21ms) to estimate force from accumulative events.

$$net_j = \sum_{l=0}^{l=3} W_{ij} \times X(n-l) + b_j \tag{3.3}$$

$$y(n) = \sum_{j} W_j f(net_j) \tag{3.4}$$

The networks are implemented in Matlab 2019 using neural networks toolbox. In this process, a cost function is defined based on the error of the estimated force from events in comparison to the measured force using a piezoresistive force sensor. To optimise the error of the network, the cost function $f(x)$ is assigned as the sum of squares of the non-linear error function $F(x)$ (Equation 3.5) where x is the input of the network. One common approach to solve such a non-linear minimisation problem is the Levenberg-Marquardt which is a combination of gradient decent and Gauss-Newton methods (Equation 3.6). In [134], it is demonstrated that the Levenberg-Marquardt optimisation technique converges faster than gradient decent with a similar accuracy. The Levenberg-Marquardt method searches for a direction in order to decrease $F(x)$ at each iteration of back-propagation (see [135] for details of calculations). In Equation 3.6, the Jacobian matrix and damping factor (non-negative scalars) are denoted as (J) and (λ) respectively and iteration number is denoted by k subscript. The damping factor is considered as 0.01 which is multiplied by an identity matrix (I) to vectorise the parameter. The Levenberg-Marquardt

method searches the directions which is given by a solution (p_k).

$$min(f(x)) = \sum_{i=1} F_i^2(x) \tag{3.5}$$

$$(J(x_k)^T J(x_k) + \lambda_k I)p_k = -J(x_k)^T F(x_k) \tag{3.6}$$

3.4.4 Gaussian Process

As mentioned in section 3.3, the triggered events and intensity changes in the scene have a logarithmic relationship. Additionally, the silicone membrane behaves non-linearly over different contact forces. Therefore, the Automatic Relevance Determination (ARD) squared exponential covariance kernel is considered to build a robust model in order to find a highly non-linear correlation between events and the contact force.

The input of GP model consists of the accumulation of positive events, accumulation of negative events and timestamps. In other words, the input has a dimension of 3×30 for three features of thirty timestamps. Although the classical GP is not trained sequentially, the timestamps information allows the model to correlate timestamps with the output. For simplicity, the GP optimisation is formulated for two variables. During the training, GP aims to minimise the error by introducing latent variables based on inputs x and desired outputs y. The inputs are the triggered events that are captured from the sensor and the outputs is the estimated contact force. Equation 3.7 presents simplified (for linear regression) version of GP where coefficients and error are denoted as α and ϵ respectively. Both α and ϵ are calculated during the training time which are used later for the inference. Equation 3.8 presents the kernel function where x_i, x_j are two inputs, σ_f is the signal standard deviation, and θ represents a logarithmic function of standard deviation of length (Equation 3.9) and standard deviation of signal (Equation 3.10)[136]. Each predictor (m) can have a different length scale (σ_m) while $m = 1, 2, \cdots, d$.

$$y = x^T \alpha + \epsilon \tag{3.7}$$

$$k(x_i, x_j|\theta) = \sigma_f^2 exp[\frac{-1}{2} \sum_{m=1}^{d} \frac{(x_{im} - x_{jm})}{\sigma_m^2}] \tag{3.8}$$

$$\theta_m = log(\sigma_m) \tag{3.9}$$

$$\theta_{d+1} = log(\sigma_f) \tag{3.10}$$

The GP model is implemented in Matlab 2019 using statistics and machine learning toolbox.

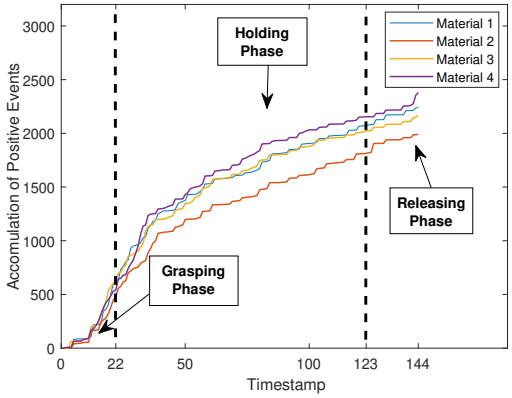
One of the most important hyper-parameters in a GP model is the length scale (σ_m) which affects the model performance significantly. In order to optimise the length scale, a Bayesian optimisation technique is performed over 10 iterations to find the best length scale. In the Matlab implementation, the Bayesian optimisation update the procedure for changing the GP model after each evaluation. Afterwards, the length scale with the best performance is selected and replaced in the kernel function to train the GP model. Finally, the same kernel function with the optimised length scale and trained hyper-parameters are used to infer the model on new inputs.

3.5 Material Classification

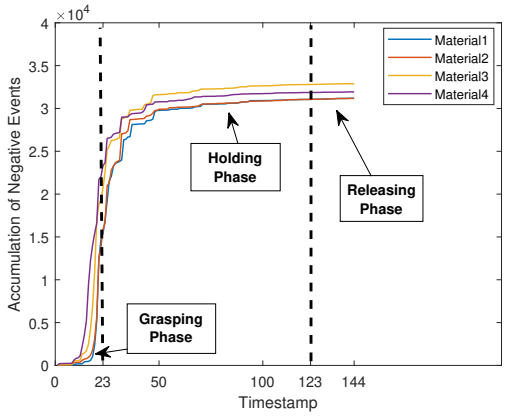
Acquiring information about the object properties such as material, friction coefficient, stiffness and weight facilitate the grasping process. In this chapter, a novel technique to classify objects' materials using DVS events from the grasping and the releasing phases is proposed. In contrast to the contact force estimation, the input for the classification model consists of two non-sequential features (accumulation of positive and negative events) without consideration of time-domain variability in the GP model. However, accumulation of events represents the time-domain features in the inputs. Figure 3.5 illustrates accumulation of events for different materials in a single grasp considering a similar range of applied force.

It can be observed that the accumulation of negative and positive events are distinguishable for different stiffness in a similar range of the applied force. The objects and the silicone membrane deform differently for each material during the grasping phase and the releasing phase. The number of positive and negative events follow different patterns for each object. Other factors such as background noise and shape of the contact area affect the number of events.

A Deep Neural Network (DNN) model is developed to classify materials considering the grasping and releasing phases. The network is designed in Matlab 2019 using neural networks toolbox with the following configurations. The network consists of k fully-connected hidden layers and n neurons in each layer. The sigmoid activation function is selected for all the layers after initial experiments. Furthermore, a soft-max function is used in the output layer to classify different materials. The Scaled Conjugate Gradient (SCG) back-propagation method is utilised to train the network. Figure 3.6 demonstrates the architecture of the proposed network for the material classification.



(a)



(b)

Figure 3.5: Accumulation of events in a single grasp for four different materials with a similar range of the contact force. (a) Accumulation of positive events. (b) Accumulation of negative events.

3.6 Experimental Setup and Data Collection

The experiments are designed to grasp four objects with different Young's modulus: (i) Foam; (ii) Rubber; (iii) Silicone; (iv) Steel. All the objects are formed in the same hexagonal shape and size (0.75×0.65×3.55cm). To increase the contrast with an opaque surface of the silicone membrane, all the objects are coloured in black. Since the objects are in the same shape and colour, the classification method only relies on the elasticity of the objects rather than the object texture or colour.

In each experiment, each object is gripped and a constant pressure is applied to hold the object for 700ms. Then, the gripper returns to the starting position to release the objects. The DVS sensor is located in a distance of 5cm from the static finger to minimise the noise and capture the changes in the contact area. A lens with 4.5mm focal length is mounted on the camera which can be adjusted regarding the size of the objects. The linear horizontal field of view of the

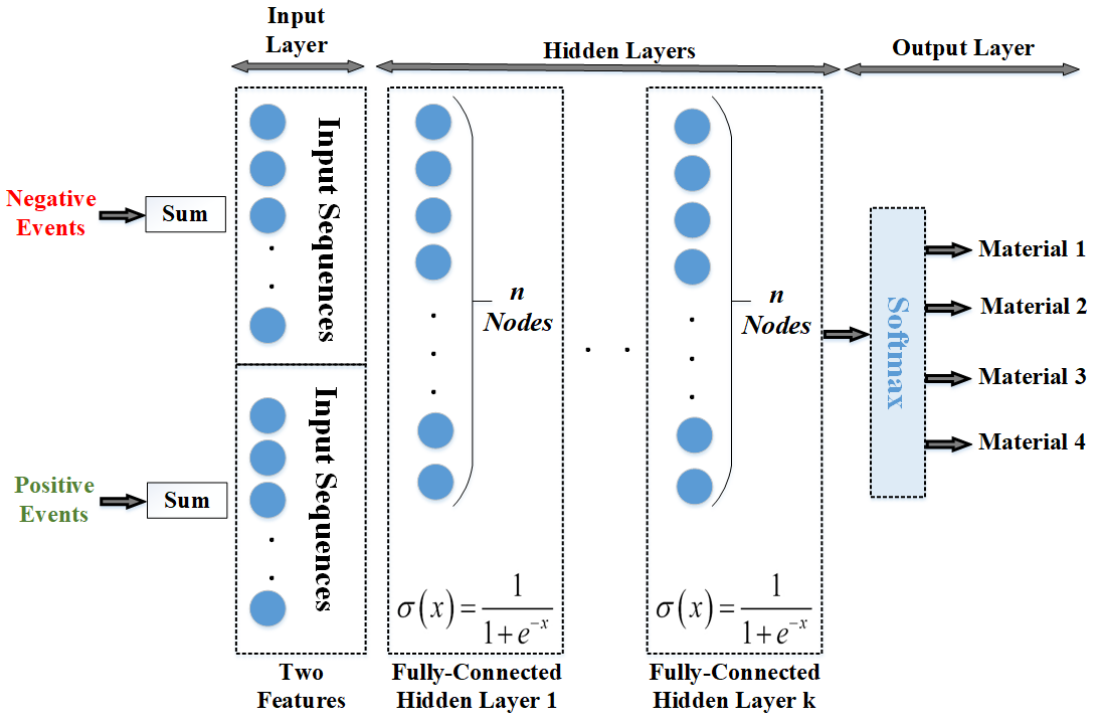


Figure 3.6: A DNN model for material classification

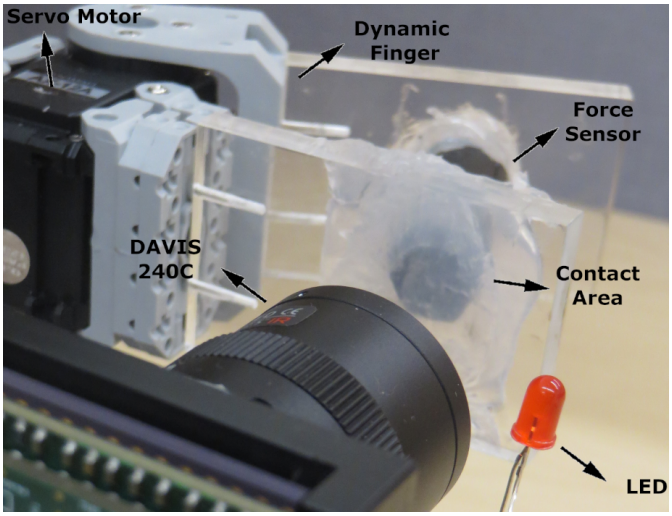
lens corresponds to 9.8cm in 10cm distance. In this setup, each pixel of the scene corresponds to 0.04mm^2 area on the silicone surface. On the dynamic finger, a piezoresistive force sensor (FlexiForce-A201) is located to measure the contact force.

The gripper composes of one dynamic and one static plane. The dynamic plane is controlled by a servo motor (AX-12A Dynamixel) using a micro-controller (Arduino) to control the gripper acceleration and position. The force is applied to the object by the dynamic finger with an angle of 15° with respect to the z-axis. Figure 3.7(a) demonstrates the experimental setup: The DVS observes the contact area through a static finger of the gripper. Figure 3.7(b) illustrates the contact area from the view of the DVS.

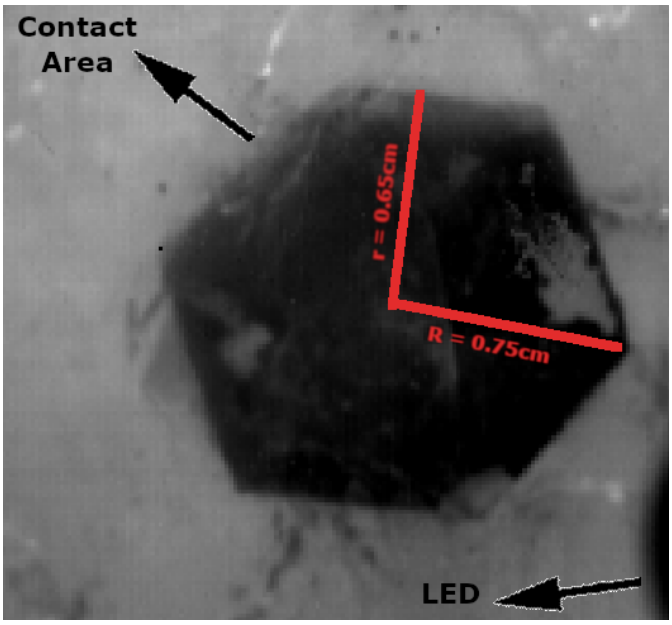
3.6.1 Force Sensor and Synchronization

A piezoresistive force sensor of FlexiForce A201 type is employed as a tactile sensor to validate the proposed event-based sensor. The force sensor has a response time $< 5\mu$s, percentage error $\pm <$ 3%, hysteresis <4.5% of full scale and is adjusted to measure forces from 0N to 111N. Moreover, experiments are performed for a range from 0N to 3.7N. This range is selected based on the saturation of silicone deformation. The force sensor is covered by a silicone layer in order to mimic the same friction coefficient on both sides of the objects.

A Light-Emitting Diode (LED) is used to synchronise the DVS camera and the force sensor.



(a)



(b)

Figure 3.7: (a) The Experimental setup includes the piezoresistive force sensors, servo motor, and two fingertips. (b) The image of the contact area from the DVS point of view.

In each experiment, the LED is turned on and after few milliseconds off prior to the grasp. When the LED is turned off, the time is recorded by the micro controller to start recording the force measurements. This time is also detected by the DVS by finding a significant spike of negative events (LED OFF) in the scene. Afterwards, the artificial frames are constructed by accumulation of positive and negative events during a 7ms window. Each experiment is divided into the grasping phase (from the 1^{st} frame to 22^{nd} frame), the holding phase (from 23^{rd} frame to 122^{nd}) and the releasing phase (from 123^{rd} frame to 144^{th} frame).

The measured force varies significantly due to the vibration and movement of the dynamic finger. A median filter is applied to smoothen the force values and filter the noise. Figure 3.8

illustrates the distribution of force across 48 experiments, captured along all 144 timestamps.

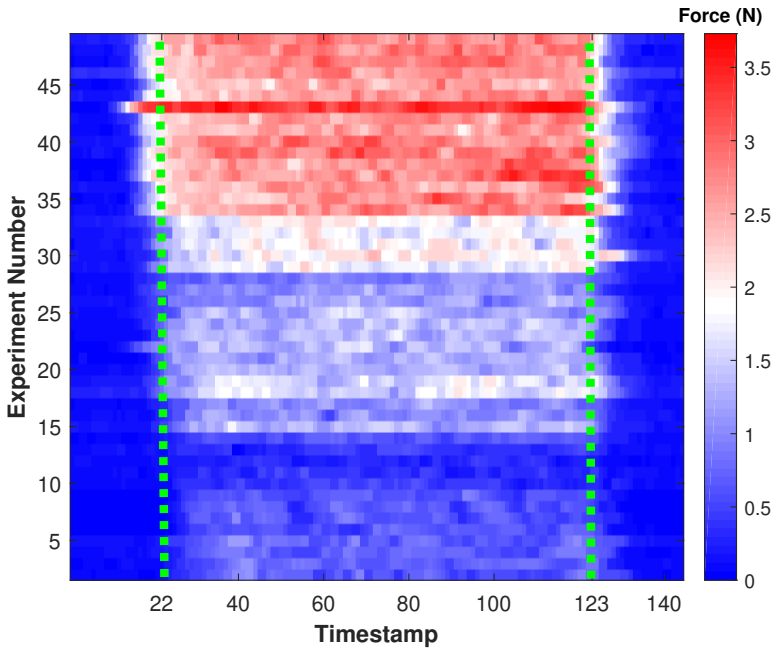


Figure 3.8: Each row represents an experiment along 144 timestamps while the colour indicates the contact force at each point. The dotted green lines show the signal clipping boundaries for the grasping, holding and releasing phases

As shown in Figure 3.8, the contact force decreases dramatically during the releasing phase. Since the force sensor is mounted on the dynamic plane and it is covered by a silicone layer, the measured contact force is small but non-zero.

3.7 Results and Discussion

This section presents the results of the proposed sensor which are validated against the piezoresistive sensor. In section 3.7.1, force estimation results are shown for the two proposed machine learning models: TDNN and GP. Section 3.7.2, presents results for the proposed DNN model that classify materials in a grasp.

3.7.1 Force Estimation

This section presents and analyses the results of the proposed TDNN and GP for force estimation. The most common approach to evaluate a machine learning method is to partition the data into training and test subsets. The model design and hyper-parameters are tuned to achieve the highest performance on the test set. The training set is given to the machine learning method to find appropriate hyper-parameters in order to minimise the error. The test subset (unseen

data) is not involved in the training process. A significant disadvantage of this approach is that researchers changes the model design and hyper-parameters based on the assessment on the test subset. Therefore, the test subset is indirectly involved in the design of the method which makes a bias in this process.

Another approach is to divide the dataset into three different partitions (training, validation and test). The validation set assists the training process to stop when the network accuracy does not improve on the validation-set. This method reduces the time of the training process and prevents the over-fitting of the network. Afterwards, the hyper-parameters are optimised on the validation set. An appropriate machine learning model and kernels can be selected by considering the method performance on the validation set. Finally, the test subset is only used to report the performance of the network rather than finding the optimal model and hyper-parameters.

In this chapter, the data is divided into three subsets: 87.5% for training (forty-two experiments), 10.4% for validation (five experiments) and 2.1% for test (one experiment). The five experiments in the validation set are selected randomly from three different ranges of the contact force to make sure that all possible values of the applied force are covered. Furthermore, an exhaustive leave-one-out cross-validation method is deployed to test each experiment individually over 48 folds. The leave-one-out method provides a comprehensive evaluation by testing the models on all of the experiments individually.

The TDNN error is calculated over all the folds (48 folds) and the average of Mean Squared Error (MSE) is calculated to compare different network architectures. To find the optimal architecture, the number of neurons and hidden layers are varied. All the network weights are initialised randomly and biases are set to zero at the first place. The networks are trained in parallel on a CPU with double precision (Corei7-8700 6cores) using the MATLAB neural network toolbox. Table 3.1 demonstrates the average MSE over all folds for different number of hidden layers (k) and neurons (n). The lowest validation error (0.15N) is achieved through a

k/n	n=5	n=10	n=15	n=20	n=25	n=30	n=35	n=40
k=1	0.23	0.19	0.18	0.19	0.20	0.18	0.16	0.16
k=2	0.21	0.19	0.16	0.17	0.17	0.17	0.17	0.17
k=3	0.19	0.20	0.18	0.18	0.16	0.17	0.17	0.17
k=4	0.20	0.19	0.16	0.17	0.18	0.17	0.17	**0.15**
k=5	0.20	0.16	0.20	0.18	0.18	0.16	0.17	0.16

Table 3.1: Mean Squared Error of the estimated force(N) on the validation set where the lowest error is highlighted in red.

network with 4 hidden layers and 40 nodes. Since the validation experiments are chosen from three different ranges of the contact force, it is expected to achieve a generalised model for the force estimation. Note that choosing different experiments for the validation partition changes

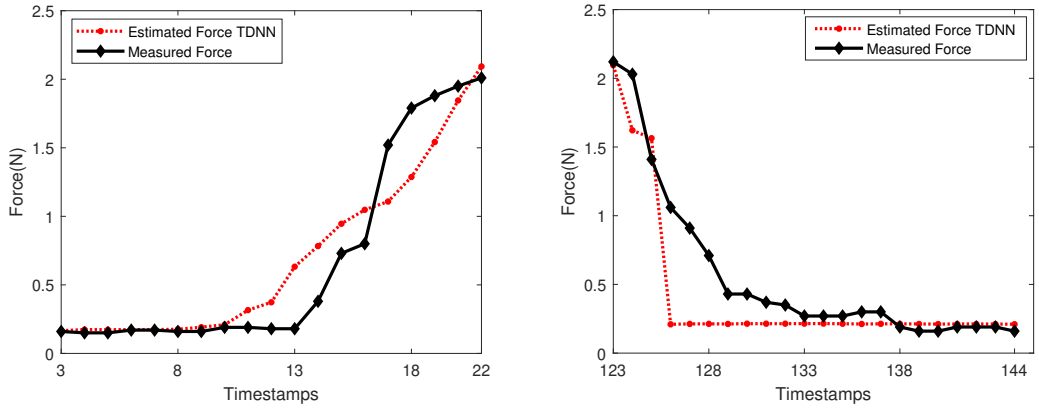
the performance of the network. Table 3.2 presents the average MSE over 48 folds for the sequences of the unseen experiments. The average MSE is highlighted for the proposed network

Table 3.2: Mean Squared Error of the estimated force(N) on the test set where the error of the proposed network architecture is illustrated in red.

k/n	n=5	n=10	n=15	n=20	n=25	n=30	n=35	n=40
k=1	0.16	0.24	0.15	0.22	0.19	0.16	0.24	0.15
k=2	0.17	0.16	0.16	0.16	0.17	0.18	0.15	0.35
k=3	0.18	0.18	0.16	0.17	0.16	0.16	0.16	0.16
k=4	0.18	0.17	0.16	0.16	0.16	0.17	0.19	**0.16**
k=5	0.18	0.16	0.24	0.17	0.20	0.15	0.24	0.18

architecture which is the second best accuracy overall. Similar performance of the network for both validation and test partitions indicates a good generalisation of the network.

Figure 3.9 illustrates the multi-layer TDNN response (red) and the measured force (ground truth) for an experiment tested on unseen data considering leave-one-out cross-validation method. As it can be observed, the estimated force follows the measured force pattern with a high accuracy during the grasping phase where the estimated force drops to a steady level. In the beginning of the releasing phase, the object loses all of the contact area with the fingertip which leads to a significant spike in number of triggered positive events. After this moment, a slight number of events are fired which indicates environment noise. Therefore, the network recognises the frames that the object is not in contact with the fingertip and it remains steady. Since the force sensor is mounted on the dynamic plane of the gripper, the measured force is affected by noise due to the motion of the gripper in the releasing phase. Moreover, the force sensor hysteresis adds a further delay to the measured force over the time. Consequently, the amount of measured force is decreasing slower over the time rather than a sharp drop at the first frame of the releasing phase. To evaluate the proposed GP model, the same folds as TDNN are considered to allow



(a) (b)

Figure 3.9: Measured force and estimated force for the TDNN on the unseen experiment during the grasping phase (a) and the releasing phase (b).

the comparison of models. The Bayesian optimization is performed on each fold individually over ten iterations to tune the hyper-parameters. Figure 3.10 illustrates the estimated force by the GP model for the unseen experiment in one of the folds. The average MSE of 0.17N is

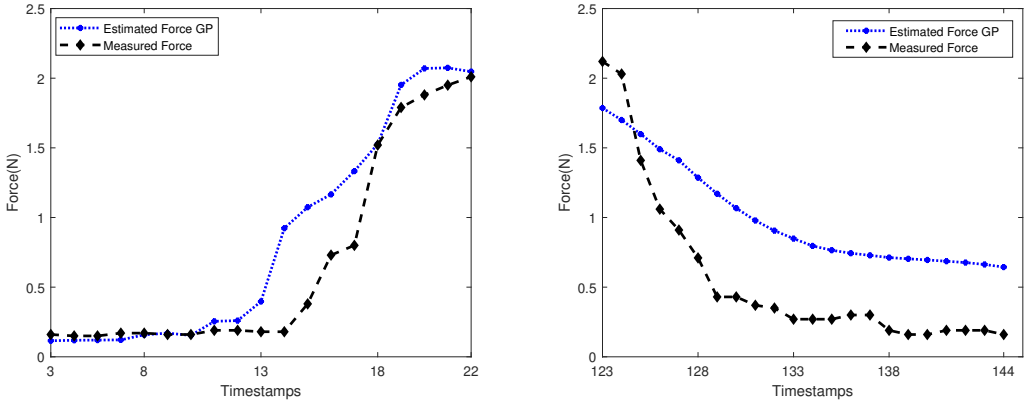


Figure 3.10: Measured force and estimated force by GP model on the unseen experiment during the grasping phase (a) and the releasing phase (b).

achieved through the time-series GP method. The response of this technique appears to be able to estimate the force in the grasping phase with a high accuracy. In the releasing phase, the GP response decreases with a slight slope compare to the measured force in this selected fold. Since the number of triggered events are close to zero in the releasing phase, the GP method learns to estimate the measured force by considering the force values as a function of time. Figure 3.11 illustrates the averaged MSE and standard deviation for the estimated force on the all folds at each timestamp for both TDNN and GP.

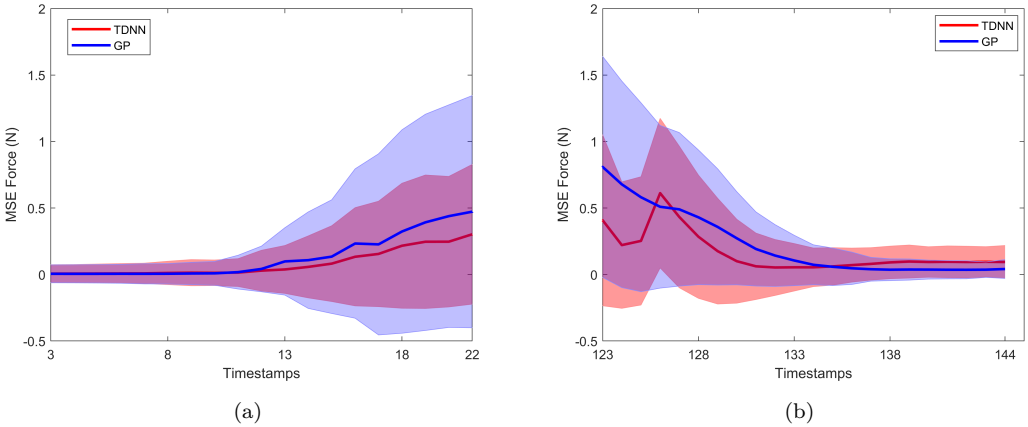


Figure 3.11: Average MSE of the estimated force are presented by red and blue lines at each timestamp during the grasping phase (a) and the releasing phase (b). The standard deviation of MSE is presented by a highlighted area over the average of MSE with boundaries of +STD and -STD.

The proposed TDNN achieved slightly higher accuracy than the GP model. The delay nodes in TDNN enable the modelling of temporal coherence of the sequences through a time window. As regards to Figure 3.11, both TDNN and GP methods identify the start of the the grasping and the releasing phases faster than the tactile sensor due to the force sensor hysteresis and experimental setup. The estimated force by TDNN drops rapidly to a low steady level, indicating the low latency of this method. Interestingly, this phenomenon is evident through most of the experiments. Figure 3.12 illustrates the same behaviour of the TDNN in a different fold.

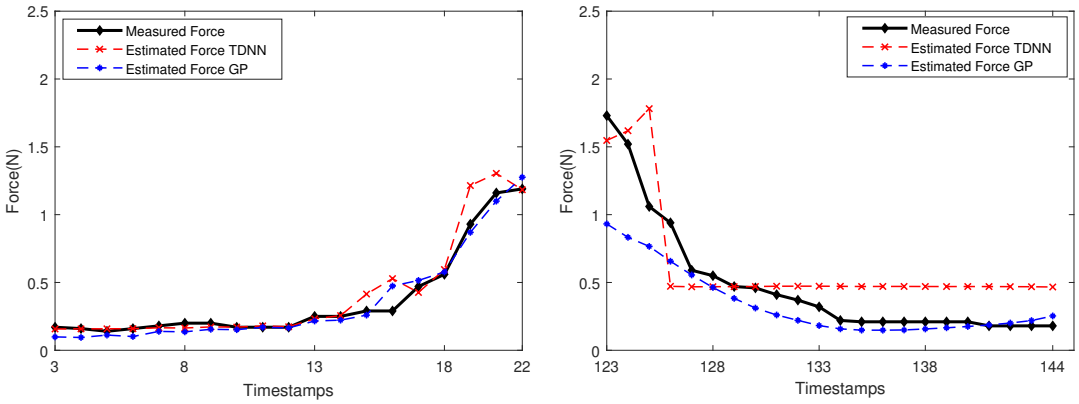


Figure 3.12: Responses of the estimated force and the measured force (groundtruth) considering two different folds.

Both GP and TDNN are well-known machine learning techniques with a low computational cost to model a highly non-linear relationship of inputs and outputs. Since TDNN has multiple delay nodes, the short-term dependencies are captured by the network during the training procedure. On the other hand, it is expected to have a higher error for GP method where the short-term dependencies are significant as GP learns the time relationships for the whole sequence.

Unlike the TDNN, the GP response shows a slight decrease during the releasing phase. Even though the triggered events in this phase is close to zero, the GP model estimates the force as a function of time. Therefore, the GP model has apparently a lower error than the TDNN where the number of the triggered events are low. However, the actual contact force must drop rapidly when the object releases. Since the force sensor is mounted on the moving gripper, the measured force includes noises until the gripper stops. Moreover, the force sensor hysteresis leads to have a considerable delay to measure the real contact force when the force varies significantly in a short amount of time.

As presented in Figure 3.11, the MSE of TDNN is lower than the GP model in the grasping

phase. Although the GP model has a better response for the last 8 timestamps whereas the measured force are not correlated with the triggered events; this part of the releasing phase represents only noise since the object is not in contact with the gripper.

The proposed neuromorphic vision-based sensor cannot be compared directly to the work of other researchers, since this method is the first work that introduces the force estimation and material classification using a neuromorphic camera (DVS). Even though marker-based and reflective sensors have shown a high accuracy, the conventional camera sampling rate limits the sensor performance. For example, in [117], a vision-based sensors with resolution of 0.05N for a maximum range of 7N is presented where each frame takes 55ms to capture. Moreover, most of the vision-based sensors evaluate the sensor accuracy in a structured environment by minimising the vibration and noise. In this chapter, the proposed sensor is evaluated in the grasping and releasing phases of a grasp which includes vibration, noise and uncertainty in measurements.

In this chapter, the proposed sensor is limited to a range of 0-3.7N for a grasp which can be adjusted by changing the silicone membrane properties. A number of experiment are performed by considering forces higher than 5N and the silicone membrane reaches the saturation point. Clearly, applying more force on the silicone membrane does not trigger further events since the silicon membrane stops to deform. This limitation is existed in all the vision-based techniques that estimate force. The range of force can be modified by considering a silicone membrane with different shape, size and properties.

3.7.2 Material Classification

The elasticity of the objects is one of the key factors in differentiating objects. The proposed classification model classifies the objects with different Young's modulus considering the grasping and releasing phases. As mentioned in section 3.6, four objects with approximately the same dimensions are tested over a wide range of forces. The number of experiments for each material is denoted in the brackets: Foam (11 sequences), Rubber (9 sequences), Silicone (14 sequences), and Steel (14 sequences).

Four experiments are chosen for the validation set and leave-one-out cross validation is implemented to evaluate the classification network accuracy. Table 3.3 represents the accuracy of the network for different numbers of hidden layers and nodes. The highest accuracy for the validation set is achieved through two models with 30 nodes. A higher number of neurons and hidden layers might lead to achieve a better result while increases the training and testing time significantly. Therefore, the network with 2 hidden layers (k=2) and 30 neurons (n) is selected

Table 3.3: Accuracy of the material classification on the validation data

k/n	n=5	n=10	n=15	n=20	n=25	n=30
k=1	80.21	76.56	83.33	84.38	85.94	83.85
k=2	70.31	81.25	84.38	85.94	86.98	**90.10**
k=3	64.06	77.60	81.77	85.42	86.98	88.54
k=4	67.19	77.60	78.65	84.38	84.90	89.06
k=5	60.42	68.75	76.56	81.25	86.46	90.10
k=6	60.94	69.79	74.48	81.77	85.42	86.46

to classify materials.

Table 3.4 illustrates the accuracy of the proposed network on unseen experiments over 48 folds.

Table 3.4: Accuracy of the material classification model on the unseen data (test set)

k/n	n=5	n=10	n=15	n=20	n=25	n=30
k=1	70.83	68.75	62.50	70.83	62.50	68.75
k=2	58.33	62.50	68.75	75.00	60.42	**79.17**
k=3	50.00	58.33	56.25	72.92	72.92	77.08
k=4	60.42	58.33	72.92	68.75	60.42	75.00
k=5	50.00	45.83	54.17	62.50	68.75	72.92
k=6	39.58	58.33	66.67	58.33	70.83	75.00

The highest accuracy (79.17%) stands for the proposed network. Figure 3.13 demonstrates the confusion matrix for these experiments considering leave-one-out cross-validation method.

As observed in Figure 3.13, the rigid material (Steel) has the highest accuracy with only one error over all folds. The classification of soft materials with a closer Young's modulus is a more challenging process. The results indicate an average accuracy of 73.3% for Foam, Rubber and Silicone.

As mentioned in chapter 2, many approaches consider multiple force sensors to classify materials and objects. To compare the proposed sensor accuracy against the piezoresistive sensor, a neural network is trained on the piezoresistive sensor readings. The force readings are considered as the input to the network to predict materials after a single grasp. The best network indicates an accuracy of 50% on the unseen data which is 29.17% lower than the accuracy of the proposed neuromorphic sensor. The main reasons of this phenomenon are the dimension of the inputs (one dimension for the force readings) and lack of history information at each timestamp. In [29], an average accuracy of 95% is reported for classifying eight surfaces of different texture and material including carpet, wood, tile and sponge. The sensor includes strain gauges and PVDF sensors embedded in the fingertip. Although the sensor achieves a high accuracy, the materials have a different texture which provides additional information for material classification.

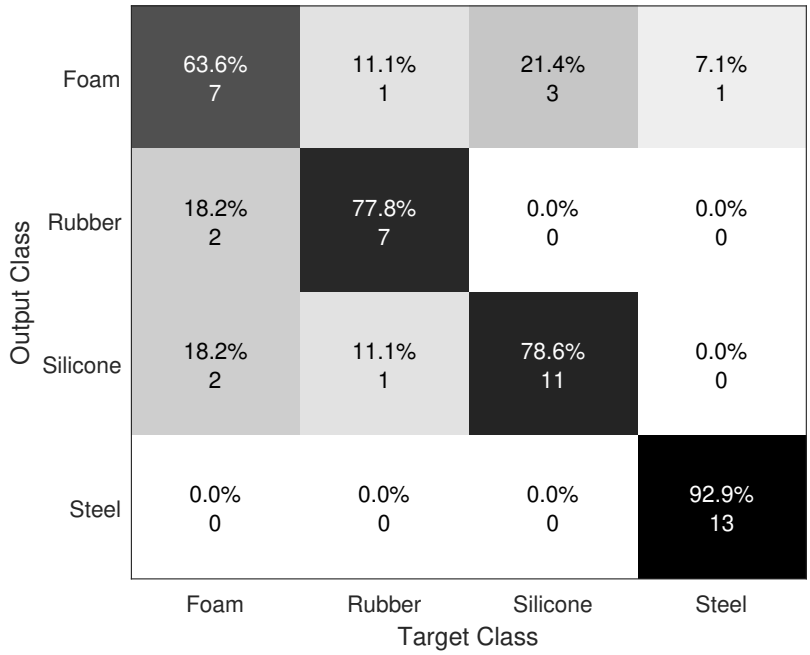


Figure 3.13: Confusion matrix for the material classification over the all folds with overall accuracy of 79.17%

.

3.8 Conclusion

In this chapter, a novel neuromorphic vision-based tactile sensor was proposed to estimate the contact force and classify four materials in a grasp. The main contributions of this chapter are:

- Design of a novel neuromorphic vision-based tactile sensor
- Development of TDNN and GP to estimate the contact force for objects with the same shape and size using spatial features
- Development of DNN to classify materials in a grasp using spatial features

A deep TDNN model with time delay of 3 nodes, four fully-connected hidden layers, and 40 neurons at each layer was developed to estimate force measurements. The TDNN estimates the contact force with the averaged MSE of 0.16N during the grasping and the releasing phases of an unseen grip. Moreover, a time-series GP model was developed which achieves the averaged MSE of 0.17N. The results indicate a promising relation between the triggered events and the contact force variation, especially if one takes into account that the source of the estimated errors may come from the hysteresis of the piezoresistive force sensor that was used to provide the ground truth. In addition, the proposed TDNN solution for the event-based force estimation seems to

have a much lower hysteresis than the piezoresistive force sensor, despite the fact that it has been trained using data from the force sensor.

Forty-eight experiments are performed on four different materials with a similar dimension and different Young's modulus. A multi-layer neural network is suggested to classify materials in a single grasp using events only. The proposed network achieves a accuracy of 79.17% on completely unseen experiments, almost 30% higher accuracy compared to the network trained on piezoresistive sensor.

The spatial information of the events are not considered in this chapter. This approach reduces the computational cost by eliminating spatial information. Accordingly, the networks model the contact force with a few features which results in a decrease in accuracy. In chapter 4, a novel technique is proposed that uses the spatial information of events for estimating the contact force.

Chapter 4

Spatio-temporal Neuromorphic Tactile Sensing

4.1 Introduction

In chapter 3, a neuromorphic vision-based tactile sensor was proposed to estimate the contact force for objects with the same size during the grasping and releasing phases using Time Delay Neural Network (TDNN), Gaussian Process (GP), and Deep Neural Networks (DNN). However, objects with different size have a variant contact area with the silicone membrane which requires a complex dynamic method to relate events and force measurements at each timestamp.

To overcome limitations of the sensor for variant object sizes, a novel technique is proposed in this chapter to estimate the contact force based on spatio-temporal event data by providing memory to the sensor using different Long-Short Term Memory (LSTM) architectures. The main contributions of this chapter are: (i) A novel neuromorphic tactile sensor for objects with a different size;(ii) Development of LSTM-based networks to estimate the contact force using spatio-temporal features.

This chapter is organized as follows. Related work is reviewed in section 4.2. The proposed neuromorphic vision-based tactile sensor is described in section 4.3. Recurrent deep learning methods for the force estimation are proposed in section 4.4. The validation process and results are discussed in section 4.5 and section 4.6 respectively. Finally, conclusions and future work are presented in section 4.7.

4.2 Related Work

In chapter 3, the vision-based tactile sensors are reviewed by considering different categories of the elastomer in the sensor and multi-modalities. In this section, a closer review in vision-based sensor is provided to highlight machine learning and image processing techniques.

Recent vision-based sensors attempt to utilise advanced computer vision and machine learning techniques to increase the sensor capabilities. For instance, a multi-task vision-based tactile sensor (GelSight) is developed to estimate the contact force and detect object slippage [137] using Convolutional Neural Networks (CNN). A pretrained CNN network (VGG-16) on ImageNet is used for transfer learning in order to estimate 3D force vector and torque over the z-axis. However, the results indicate that CNN networks cannot be generalised well on GelSight images as different contact geometries generate various features in images for similar amount of force.

In [138], Finite Element Model (FEM) is taken into account to generate groundtruth for the contact force distribution. Afterwards, an optical-flow-based tracking algorithm is considered to create a feature vector which is correlated to the 3D reconstruction of the contact force by neural networks with very high accuracy for an object with the same size and shape. Initially, optical flow features are extracted from the contact area. Afterwards, the extracted features are fed as inputs into a non-sequential deep neural network model to estimate the contact force. Although the optical flow features are utilised for training of the network, still the network is not sequential and cannot relate force features continuously in a dynamic manner. Conversely, an inverse Finite Element Model (iFEM) is considered in [121] to reconstruct the contact force distribution in three dimensions. However, the experiments are performed with a single object of known geometry and the normal contact force is limited between 3-4N.

In addition to different algorithms and techniques in the vision-based tactile sensing, other factors such as fingertip material and pins design affect the sensing performance significantly. Ward-Cherrier et al. [53] designed a wide range of bio-inspired and 3D-printed fingertips (TacTip) with various specifications to localise objects with less than 0.2mm error based on pins displacements. The experiments on different fingertips show that the pins specifications have a significant impact on the tracking algorithm and the sensor accuracy.

In other work [139], a sequential network is utilised to estimate object hardness independent from shape using GelSight. The sequences of images are captured in the loading phase and subtracted from the first frame. Then, a CNNLSTM network (spatio-temporal) is trained on sequences of subtracted images by considering the loading phase only. The dynamic network provides capability to the sensor to deal with shape-independent objects for the hardness estimation

after a complete loading phase. Even though the changes in intensity obtained by subtracting frames, still conventional cameras suffer from low sampling rate as well as low dynamic range which limits the sensor performance.

Neuromorphic vision-based tactile sensing is relatively a new research field which aims to utilise event-based cameras to acquire physical properties in the contact area. In the earliest work [130], a novel method is proposed to employ a neuromorphic camera (DVS) in order to detect incipient slippage using traditional image processing techniques. Similarly, a neuromorphic tactile sensor is proposed in [140] to detect fast phenomena such as object slippage with high temporal resolution of 500µs by accumulating events without consideration of spatial information.

Following my work in chapter 3, a considerable attention has been paid to neuromorphic vision-based tactile sensors by researchers in 2020. For example, a contact-level classifier is developed in [141] to classify objects size and the contact force range. Authors considered a similar approach to chapter 3 by accumulating the events for the classification task. Afterwards, Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) machine learning methods are implemented sequentially while both achieve approximately 89% accuracy for the object size classification.

Moreover, a feature based slip detection is proposed in [142] by using Harris detector. The spatial features such as corners and edges are extracted in each event-frame. Afterwards, the features of each frame are compared to the previous frames to validate the detection. In other work [143], a new model of TacTip sensors (NeuroTac) is proposed which combines pin-based membranes and neuromorphic vision sensors to classify objects' texture with KNN learning method. The results show that spatio-temporal coding of events improves the texture classification by more than 7% compared to spatial features only. Table 4.1 summarises specifications of vision-based tactile sensors in the literature for different applications.

Remarkable progress has been made in vision-based tactile sensors in terms of accuracy and resolution. Nevertheless, most of the aforementioned research in literature considers a machine learning approach without sequential layers for the measurements. Therefore, the sensor ignores the history of measurements to estimate the current values. CNN-based methods like [137], have shown deficiency of static network performance for objects with different geometry while sequential networks in [139] can handle different shapes. Even though silicone material provides a physical memory to the system, non-sequential networks suffer from time-related variables. In contrast, the dynamic (sequential) networks adapt the measurements based on the previous observed part of the sequences. For instance, a large and a small bolt have different contact areas with the same contact force. Therefore, the static networks cannot adopt the measurements based on CNN features only, while sequential networks take into account the object size in early

Table 4.1: Comparison of state-of-the-art vision-based tactile sensor where σ represents standard deviation). The resolution of 3D force sensors are considered for the z-axis (normal force) measurements only.

Ref	Camera	Purpose	Method	Specifications
[138]	Frame-based	Force Distribution	Optical Flow DNN	• Precise 3D force distribution • Resolution of 0.003N (MSE) • Size-dependent
[137]	Frame-based	Force Measurement	CNN Transfer Learning	• 3D force estimation • Resolution of 1.44N (MSE) • non-sequential network
[139]	Frame-based	Hardness Estimation	CNNLSTM	• Various objects shape • Only grasping phase • sequential network
[121]	Frame-based	Force Distribution	Marker Tracking	• Range of 3-4N • σ =0.322N for normal force • Spatial-temporal image processing
[130]	DVS	Incipient Slip Detection	Morphological Operations	• Latency of 44.1ms • Shape and material independent • Traditional image processing
[140]	DVS	Incipient Slip Detection	Image Analysis	• Slip detection • Event-framing over 500µs • Orientation estimation
Chapter 3 [144]	DVS	Force Estimation	TDNN and GP	• Logical latency 21ms • Resolution of 0.16N (MSE) • Various materials
This chapter [145]	DVS	Force Estimation	LSTM-Based Networks	• logical latency of 10ms • Resolution of 0.064N (MSE) • Spatio-temporal event data

stages and adopt measurements continuously.

4.3 Spatio-Temporal Force Estimation

In this chapter, a novel technique is proposed to estimate the contact force for objects with different size. The pipeline of the proposed dynamic-vision-based tactile sensing is demonstrated in Figure 4.1, where a neuromorphic camera (DVS) is employed to capture events. Firstly, sequential frames are constructed from events. Afterwards, the constructed frames are processed by Recurrent Neural Network (RNN) to estimate the contact force dynamically. Each stage of the proposed framework is discussed in the following sections.

4.3.1 Construct Frames

The output of DVS is a stream of events, each characterised by position, timestamp and polarity (x_k, y_k, τ_k, p_k) where k is a counter for events and p_k represents polarity of the pixel. Polarity is defined in a binary format which be either 0 or 1 for negative and positive polarities respectively ,as show in Figure 4.1. In Chapter 3 [144], the events are accumulated over time without consideration of position information for events to estimate the contact force. In this chapter, an event framing technique is considered with spatial information of events to create sequential frames.

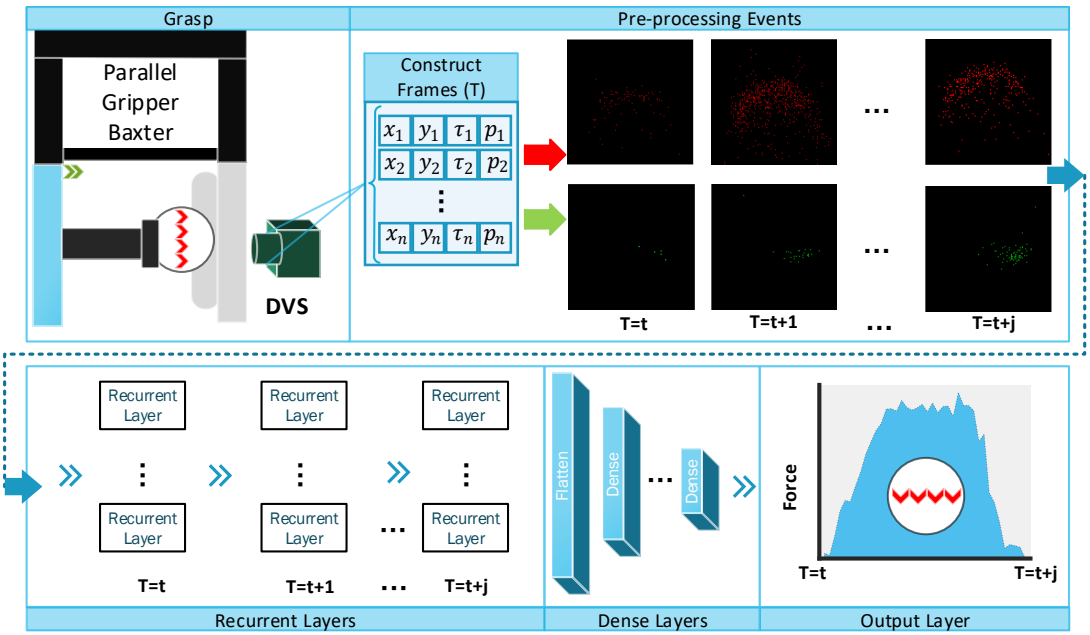


Figure 4.1: Schematic diagram of the proposed dynamic sensor for the continuous force measurements in a grasp.

To process events as a group, events are accumulated within a time window (T). A larger time window increases the number of visible events within a frame. Consequently, each frame has a considerable number of features. On the other hand, a short time window decreases the sensor latency while providing few events at each frame. The window size must be chosen by considering the application speed, the sensor latency and the object size. In this thesis, the window size is selected as 7ms and 10ms for the chapter 3 and 4-5 respectively. This range of time window provides a lower latency against the conventional cameras while making sure that enough events are triggered at each frame for the grasping task. Furthermore, decreasing the window size results in generation of redundant frames which increases the memory requirements of the system. Positive and negative polarities (p) represent a reduction or increase in the contact force respectively. Therefore, the accumulation of events are performed separately over two channels. Adapted from [71], the framing process is formulated in Equation (4.1), where X_t represents histogram of events in three dimensions for location of each pixel and polarity $p = \{0, 1\}$, Kronecker delta function (Equation 4.2) and rectangle function (Equation 4.3) are denoted as δ and Π respectively. Timestamp of each event is represented as τ_k where k indicates the event number.

$$X_t(x, y, p) = \sum_{\forall k} \Pi\Big(\frac{\tau_k}{T} - 0.5 - t\Big)\ \delta_{xx_k}\ \delta_{yy_k}\ \delta_{pp_k} \tag{4.1}$$

$$\delta_{ij} = \begin{cases} 0 & \text{if i}\neq \text{ j} \\ 1 & \text{if i=j} \end{cases} \tag{4.2}$$

$$\Pi_{(x)} = \begin{cases} 0 & \text{for } |x| > \frac{1}{2} \\ \frac{1}{2} & \text{for } |x| = \frac{1}{2} \\ 1 & \text{for } |x| < \frac{1}{2} \end{cases} \tag{4.3}$$

Each experiment is converted to a number of sequential frames (images) each of which represents intensity changes within a time window. The experiments are slightly varied in the length and empty frames are added to equalise the number of frames per experiment.

4.4 Dynamic Force Estimation

A dynamic sensor captures changes in measurements rather than their absolute values. Similarly, DVS records intensity changes and the history of each pixel is required to relate the force measurements to the triggered events at each frame. Therefore, RNNs are an appropriate method to estimate the contact force based on history of frames over time. The main advantage of RNNs is an internal state that enables the network to capture sequential dependencies between variables over time. The major problem of basic RNN is the vanishing gradient when the network fails to learn long dependencies and the gradient decent stops to converge. In a grasp, the current contact force values can be estimated by looking at the history of the contact force. Since DVS captures the dynamics of the scene, the long-term dependencies will have a direct impact on the sensor performance.To solve this problem, Long-Short Term Memory (LSTM) and Gated Recurrent Units (GRU) are proposed to control memory states [99]. In this chapter, three architectures are proposed by combining LSTM layers with convolutional and dense layers to estimate the contact force.

4.4.1 Long Short-Term Memory Units

LSTM networks have made a significant breakthrough in time-series applications such as speech recognition and action recognition. A typical LSTM unit includes input gate, forget gate and output gate which allows the network to forget unnecessary dependencies to prevent vanishing gradients. Suppose that X_t is the input (image) of the contact area to the LSTM unit and c_t is a memory cell that accumulates states at each time.

For every timestamp, input or update gates i_t will be activated and control the forget gate (f_t). Then, the forget gate decides the remaining images in the memory cell (c_t). Afterwards, the output gate (o_t) controls the use of images from the final state of LSTM (h_t). The forget gate and final state of the LSTM cell is initialised as zero for the first step. The controlling process of multiple gates allows the LSTM unit to be robust against the vanishing gradient problem to capture dependencies between the contact force and constructed frames. The main equations of an LSTM unit are presented in Equation.(4.4) where sig is the activation function and Hadamard product is denoted as $\circ$. Two matrices for inputs weights and recurrent connections are presented as W and U respectively. The initial value of c and h are defined as zero for the first step.

$$f_t = sig(W_f X_t + U_f h_{t-1} + b_f) \tag{4.4a}$$

$$i_t = sig(W_i X_t + U_i h_{t-1} + b_i) \tag{4.4b}$$

$$o_t = sig(W_0 X_t + U_0 h_{t-1} + b_o) \tag{4.4c}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ sig(W_c X_t + U_c h_{t-1} + b_c) \tag{4.4d}$$

$$h_t = o_t \circ sig(c_t) \tag{4.4e}$$

The subscripts of the weight matrices indicate the input gate i, the forget gate f, the memory cell c or the output gate o while biases for each gate are presented by the gate subscript b. Often, LSTM gate's activation functions are considered to be either sigmoid or hyperbolic tangent function. All gates, including forget Equation (4.4a), input Equation (4.4b), output gates Equation (4.4c) and memory cell Equation (4.4d) are dot-products of the weights matrices with hidden states. In fact, LSTM cells operate on vectors and disregard spatial information of inputs and hidden states.

A single LSTM unit is sufficient for basic applications whereas the input and output relationships are not highly non-linear. However, applications with a high degree of non-linearity require further learnable parameters and multiple hidden layers to model behaviour of variables. Stacking LSTM units adds further learnable parameters and enables the network to model very complex relationships between the triggered events and the contact force with consideration of the silicone deformation. The optimal number of hidden layers and LSTM cells requires to be tuned by trial and error.

4.4.2 Convolutional Long-Short Term Memory Layers

Convolutional LSTM (ConvLSTM) networks are relatively new modified versions of LSTMs that can capture spatial-temporal dependencies. ConvLSTM is proposed in [146] to forecast weather conditions where the spatial-temporal dependencies are significantly important. The main difference of ConvLSTM layers is to replace multiplication operations with convolution denoted as (*) for controlling the gates as shown in Equation (4.5).

$$f_t = sig(W_f * X_t + U_f * h_{t-1} + b_f) \tag{4.5a}$$

$$i_t = sig(W_i * X_t + U_i * h_{t-1} + b_i) \tag{4.5b}$$

$$o_t = sig(W_0 * X_t + U_0 * h_{t-1} + b_0) \tag{4.5c}$$

$$c_t = f_t \circ c_{t-1} + i_t \circ sig(W_c * X_t + U_c * h_{t-1} + b_c) \tag{4.5d}$$

$$h_t = o_t \circ sig(c_t) \tag{4.5e}$$

Opposed to LSTM layers, ConvLSTM layers maintain both spatial and temporal information of each frame. Due to the non-linearity of silicone material, spatial and temporal event data need to be taken into account for estimation of the contact force during different phases.

4.4.3 Convolutional Layers with LSTM

Convolutional Neural Networks (CNNs) are designed to extract both spatial and temporal event data in the image by applying convolution operations of filters in a certain window size. CNN networks have achieved significant success in different applications such as AlexNet [88] for image classification task which made CNNs a gold standard in modern computer vision. In addition, time-dependent applications such as video classification consider an architecture with combination of CNNs and RNNs architectures [147]. As triggered events are accumulated in a frame, CNNs are used to extract features of each frame to correlate accumulation of events with force values. Afterwards, the output of CNN layers requires to be correlated with force over time. Therefore, LSTM layers are considered after CNN layers to provide temporal memory for the extracted features (CNNLSTM). In the final layers, the output of LSTM is connected to the dense layers to estimate the contact force dynamically for objects with different sizes.

The main difference between CNNLSTM and ConvLSTM architectures is the order of per-

forming convolution operations on the constructed frames. In CNNLSTM, convolution operation applies on the frames to extract features which is followed by LSTM units to model extracted features temporally over time. However, ConvLSTM operates convolution inside the LSTM gates which maintains both spatial and temporal information of the constructed frames.

4.5 Experiments

This section presents detailed information of the experimental setup, the pre-processing stage and the networks' implementation.

4.5.1 Experimental Setup

Similar to chapter 3, the experimental setup includes an ATI F/T sensor (Nano17), a DVS sensor, and a Transparent 3D printed plane (static plane) for the Baxter robot which is covered by a silicone layer as shown in Figure 4.2.

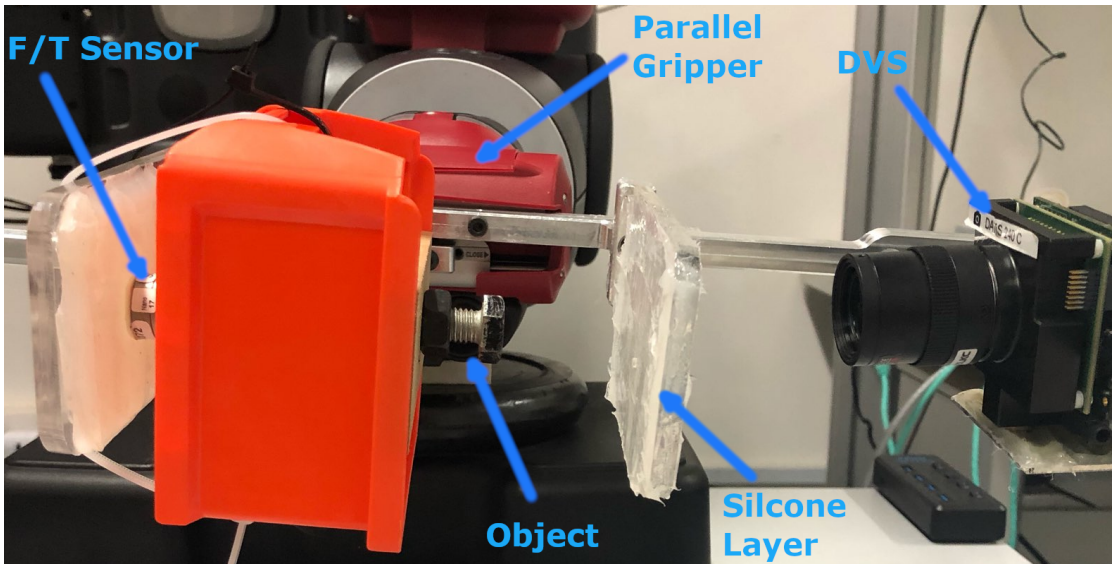


Figure 4.2: Experimental setup to perform experiments using Force/Torque (F/T) sensor and DVS on the Baxter parallel gripper.

The primary differences of experimental setups between this chapter and chapter 3 are summarised in Table 4.2. The object is centered to the the plane to grasp different objects with the same center of contact area on the silicone.

For each experiment, the gripper is calibrated first and the same process of closing and opening the grippers are followed. Although the experiments are performed with the same configuration,

Table 4.2: Comparison of experimental setup parameters between this chapter and chapter 3

Parameters	**Chapter 3**	**This Chapter**
Object Material	4 materials	1 material
Object Size	1 size	3 sizes
Force Sensor	FlexiForce-A201	ATI Nano17
Maximum Force	3.7N	3.12N
Maximum Length	1008ms	410ms
Time Window	7ms	10ms
Gripper	PhantomX Parallel	Baxter

the contact force values and experiments duration are slightly varied due to the silicone elasticity, controller delay, and measurement uncertainty. As the sensor estimates the contact force continuously, each experiment is divided into the grasping, holding and releasing phases as can be seen by a visual summary of the experiments in (Figure 4.3). Thirty-five experiments are performed on three bolts with size of 8mm,12mm and 16mm.

In each experiment, the contact force starts from zero and reaches the maximum of 3.12N during the holding phase. The releasing phase is highly non-uniform across experiments due to the elasticity of the silicone membrane.

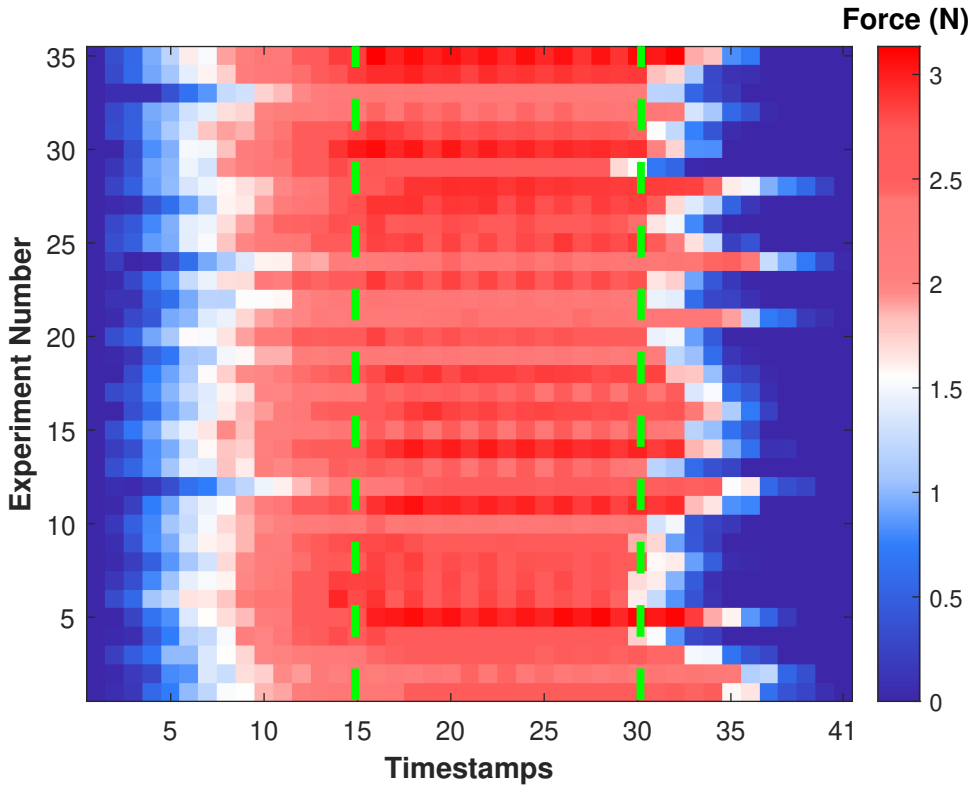


Figure 4.3: Each row shows one experiment over time. The colour represents the force values (N) from low contact force (blue), medium contact force (white), and high contact force (red). The green dotted lines differentiate between the grasping, holding and releasing phases.

4.5.2 Pre-processing

The framing process accumulates events in two different channels, as described in section 4.3. The time window is selected as T=10ms, which ensures that sufficient number of events are accumulated in frames. Furthermore, the frames are cropped based on the largest contact object contact area from 240×180 to 115×115 to reduce the memory requirements. The thresholds of positive and negative polarity are tuned to reduce noise levels. The maximum number of accumulated positive and negative polarity events across all experiments are considered in all experiments to avoid saturation in the constructed frames, i.e. their pixel values do not exceed the maximum 8-bit range. Thus, a weight function is applied to normalise the value of each pixel in two channels. Since the implementation of RNN requires a fixed length of sequences, the latter is derived by the longest experiment (410ms, 41 timestamps) and other experiments are zero-padded to this length.

4.5.3 Neural Network Implementation

The proposed networks are tuned based on trial and error considering various hyper-parameters including number of hidden layers, filters and dropout rate. The networks are designed in python with Keras framework [148] and trained on a NVIDIA GTX 1080 Graphical Processing Unit (GPU). LSTM cells are initialised with random orthogonal matrices which improves the robustness of LSTM layers to prevent vanishing gradient [149]. The loss function is considered as the Mean Squared Error (MSE) which penalise the error for higher force values.

Adam optimiser has shown efficient converge to optimise neurons' weight efficiently in terms of memory and speed [150]. Furthermore, Adam is appropriate for sparse and noisy data which is considered in this chapter with learning rate of 0.001. Also, Adam coefficients of moving average (beta1 and beta2) are set to 0.9 and 0.999 respectively.

The number of layers and network hyper-parameters are chosen by trial and error. The frames represent events over a short period of time which reduce the number of features in each frame. Additionally, each event in the frame is related to the contact force directly. As the objects have a flat surface, the number of filters are retained minimal to reduce the number of hyper-parameters. Therefore, convolutional layers are designed with three filters in size of 3×3 and zero padding in each layer. Furthermore, drop out layers are used in the first three layers with rate of 0.4 to prevent over-fitting on the training set.

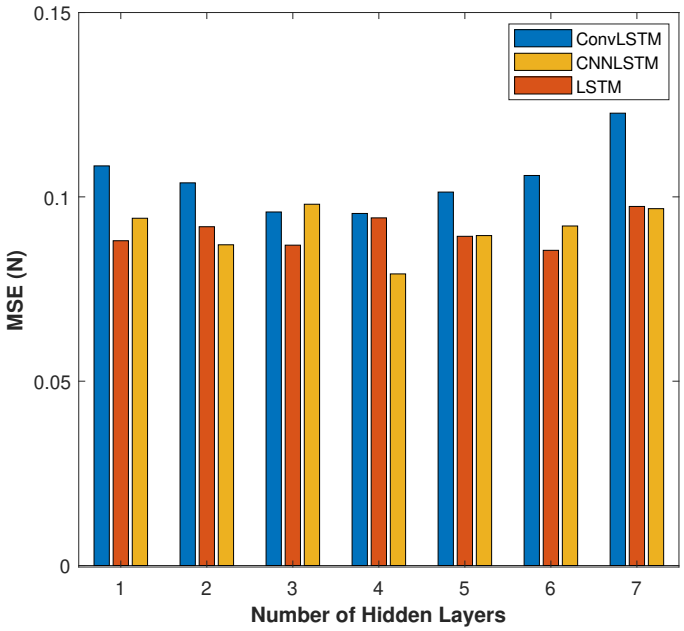
The data is split into three sets of training (31 sequences), validation (3 sequences), and test

(1 sequence) sets. The data collection process is time-consuming and costly since the objects are required to be placed manually in the setup. Therefore, a limited number of experiments are performed to validate the proposed sensor. To generalise this solution, other techniques such as data augmentation (Chapter 5) and simulations can be considered to increase the volume of the data. Since the experiments are limited in this chapter, leave-one-out cross-validation method is implemented to evaluate the sensor performance. This cross-validation technique is an exhaustive case of K-fold technique where K is equal to the total number of experiment (35 folds in this chapter). Consequently, 35 different models are trained by leaving one example for the test in each fold for the testing purpose. The final results are achieved by taking an average of the networks accuracy for all folds. In addition, the training process is controlled by early stopping technique by considering 20 iterations after the last improvement of network accuracy on the validation set. This process prevents the over-fitting problem by stopping the training process. To select the validation set, a random experiment is picked up from each size to ensure all object sizes are included. Therefore, the evaluation metric on the validation set is representative for all the object sizes to prevent a bias towards a specific object size during the training process.

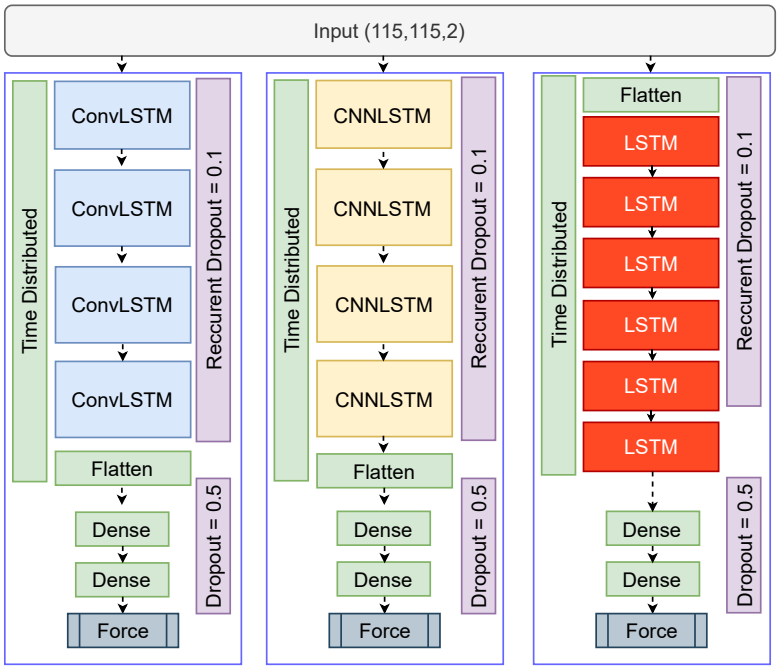
4.6 Results and Discussion

The validation set is only used to tune hyper-parameters and select the optimal architecture. Specifically, the average of Mean Squared Error (MSE) over all folds is calculated on validation set to select the best hyper-parameters for each architecture. Figure 4.4(a) illustrates the average MSE over all folds for the validation set considering a range of 1 to 7 hidden layers. The minimum MSE is achieved with 4 hidden layers for ConvLSTM and CNNLSTM while LSTM network reaches the best performance with 6 hidden layers excluding the dense layers. The models' run-time is approximately 0.12 sec which can be further reduced with the speed optimisation frameworks such as TensorRT. In average, the training time for a model takes 45 mins considering ConvLSTM architecture and 35 experiments. The training time depends on GPU model, number of experiments, and deep learning framework (Keras). Figure 4.4(b) presents the networks' architectures and configurations including dropout layers for the three dynamic methods.

After selecting a model with the lowest MSE for each architecture, average of MSE on a test set is calculated for evaluation purposes. Similar to [138], only non-zero measurements are considered to provide a realistic assessment of the sensor. Finally, the same experiments are used to train a TDNN network to compare the proposed methods with the previous chapter 3).



(a)



(b)

Figure 4.4: (a) Average MSE of Validation over all folds by varying number of hidden layers from 1 to 7 for ConvLSTM (Blue), LSTM(Red) and CNNLSTM (Orange). (b) Network architectures and configurations for ConvLSTM, CNNLSTM and LSTM.

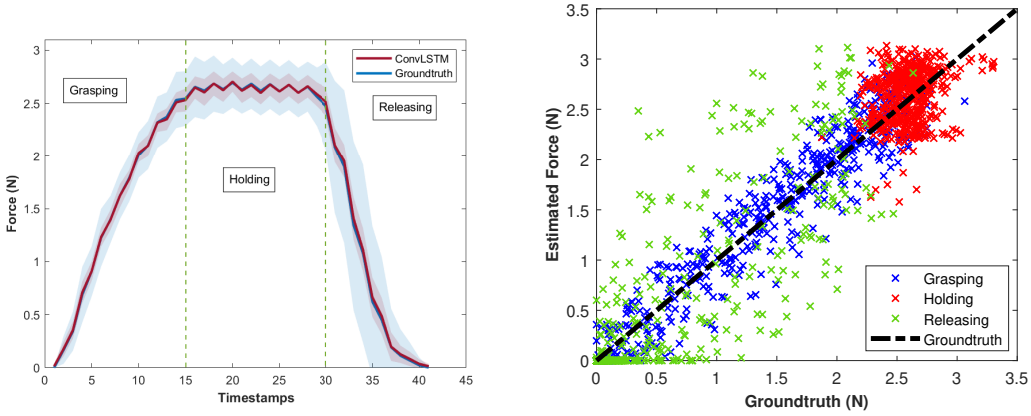
Table4.3 presents the average of Mean Absolute Error (MAE) and Mean Squared Error (MSE) over all the folds where the standard deviation is denoted as σ.

Although ConvLSTM validation error is higher than other models, this architecture generalises better for the test set and achieves the highest accuracy considering all the phases. Figure

Table 4.3: Average error of the estimated force and standard deviation (σ) for the test set.

Network/Errors	MAE(σ)	MSE(σ)
TDNN	0.398(0.410)	0.345(0.713)
LSTM	0.301(0.234)	0.160(0.261)
CNNLSTM	0.291(0.234)	0.157(0.259)
ConvLSTM	**0.278(0.225)**	**0.145(0.237)**

4.5(a) demonstrates the average of the estimated force and groundtruth over all folds during the grasping, holding and releasing phases Figure 4.5(b) demonstrates the estimated force values against the groundtruth for different phases.



(a) (b)

Figure 4.5: (a) Average of the estimated force and groundtruth for three phases over all folds. The highlighted area presents the standard deviation of values. (b) The scatter plot for estimated force is presented in the grasping (blue), holding (red) and releasing (green) phases against the groundtruth (black line).

The deformation of silicone material in different phases is highly non-linear due to the changes in the contact force and object size. Most of studies in literature (see Table 4.1) consider only the loading phase to eliminate the impact of silicone elasticity from the measurement. Table 4.4 presents the average of MSE and standard deviation of error for each phase considering all folds.

Table 4.4: Average of MSE (N) and standard deviation (σ) of the estimated force over all folds for unseen experiments.

Phase	Grasping		Holding		Releasing	
Methods	MSE	σ	MSE	σ	MSE	σ
TDNN	0.309	0.572	0.190	0.426	0.490	0.425
LSTM	0.065	0.053	0.092	0.065	0.537	0.407
CNNLSTM	**0.063**	**0.051**	0.088	0.064	0.527	0.386
ConvLSTM	0.064	0.055	**0.082**	**0.063**	**0.485**	**0.372**

The results indicate that the proposed approach for force estimation is very promising for the grasping phase where all three LSTM-based networks achieve similar results. Due to the

Table 4.5: Comparison of Dynamic Time Warping Distance (D_W) and Bhattacharyya Distance (D_B) for three different networks' architectures.

Phase	Grasping		Holding		Releasing	
Methods	D_B	D_W	D_B	D_W	D_B	D_W
LSTM	0.102	1.711	2.278	3.811	0.622	2.282
CNNLSTM	**0.009**	**1.642**	2.483	3.588	0.061	**2.204**
ConvLSTM	0.009	1.678	**1.175**	**3.258**	**0.055**	2.346

elasticity of the silicone membrane and vibrations, errors tend to be more significant during the holding and releasing phases. Also, the difference between accuracy of different network architectures increases continuously towards the end of releasing phase.

Furthermore, non-linearity of the silicone behaviour creates a variable time lag between the F/T sensor and the proposed sensor. In order to investigate this problem, we perform Dynamic Time Warping (DTW) and calculate the distance between the estimated force and groundtruth in different phases. To measure similarity of distributions between the estimated force and the F/T sensor, Bhattacharyya distance is considered for each phase of the grasp. Table 4.5 presents average of Bhattacharyya distance (D_B) and DTW distance (D_W) over all folds for each phase of the grasp. Nevertheless, a slight time lag in the estimation leads to have a similar MSE in the holding phase for both CNNLSTM and CONVLSTM.

As presented in Table 4.5, both CNNLSTM and ConvLSTM achieve similar results considering DTW and Bhattacharyya distance in the grasping phase. However, ConvLSTM achieves significantly lower D_B during the holding phase. LSTM cells in CNNLSTM architecture perform on a vector which results in loss of spatial-temporal information while Convolutional LSTM (ConvLSTM) networks are more robust by keeping the input dimension inside LSTM layers. The low values of D_W and D_B for ConvLSTM in the holding phase indicate that it follows the contact force variations, caused by vibrations, better than other networks.

In Figure 4.6, the estimated force and groundtruth for different sizes of bolts are illustrated. The ConvLSTM network achieves the highest accuracy for small objects while CNNLSTM performs better predictions for medium and large objects. The main reasons for various errors are limitation of data, random selection of experiments for the validation set, and non-linearity of silicone behaviour considering a larger contact area. The experimental setup is uncontrolled to evaluate the proposed method for practical applications. The experiment lengths are varied slightly for different objects due to the elasticity of the silicone membrane. Therefore, a slight difference in the contact force measurements distribution is recognisable in the experiments which affect accuracy of the network for different objects.

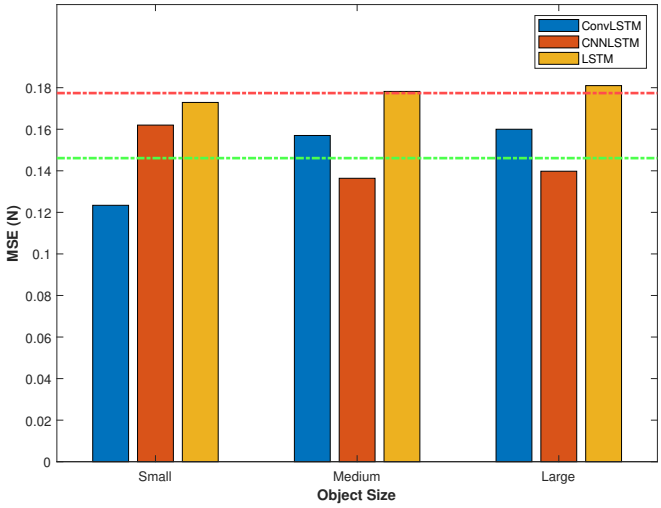


Figure 4.6: The green line shows average of MSE for ConvLSTM and CNNLSTM networks approximately. The red line illustrates the average of MSE for LSTM network.

As demonstrated in Table 4.1, the majority of the vision-based tactile sensors utilise static (without memory) DNNs without considering history of events in the contact area. In [138], Deep Neural Network (DNN) is employed with optical flow inputs to estimate the contact force. The reported results (RMSE=3mN) are obtained in a controlled environment, where force values are obtained after the stabilization of a single object (intender). When a static DNN, such as the CNN-based transfer learning approach in [137] is applied in a variety of objects iwth different sizes during the grasping phase, errors are much higher (RMSE=1.2N).

In contrast, we proposed a dynamic approach to estimate the contact force continuously for objects with different sizes. Our approach provides a memory to the sensor in order to learn objects geometry at early stage of the grasping phase and adopt the contact force estimation during different phases. In the grasping phase, the proposed sensor achieves MSE=0.064N, comparable to state-of-the-art VBM sensors (Table 4.1), for objects with different sizes. In addition, MSE of 0.082N is achieved during the holding phase where the object vibration is inevitable. It should be noted that the holding phase has not been considered by other aforementioned VBM methods.

Furthermore, we trained a new network (ConvLSTM 4 hidden layers) with small and large size objects and validated it on the experiments with the medium object. Twenty-eight experiments of large and small objects are used for training while twelve experiments of the medium size are considered for validation and testing (six for validation and six for the test set). The results indicate MSE of 0.159N and MAE of 0.385N on the test set during the all three phases. As expected, the error is higher but still acceptable when the sensor is used on objects with sizes

different from the ones used for training.

4.7 Conclusion

In this chapter, a novel methodology is proposed to estimate the contact force dynamically considering spatio-temporal event data. The main contribution of this chapter are: (i) Development of a novel neuromorphic visio-based tactile sensor to estimate the contact force for objects with a different size ;(ii) Implementation of LSTM-based networks to estimate the contact force based on spatio-temporal event data for objects with different sizes.

A novel dynamic approach is proposed to estimate the contact force for size-variant objects. The main challenge of force estimation for size-variant objects is different contact area geometry under similar amount of applied force. It is demonstrated that LSTM-based networks learn to relate the contact area size to the corresponding contact force over time.

Three LSTM-based networks are developed and implemented to estimate the contact force based on history of changes in every pixel considering both spatial and temporal information. The proposed sensor is validated on Baxter robot for three bolts with different sizes. ConvLSTM achieved the best results, specifically MSE=0.064N for estimating the contact force in the grasping phase and MSE=0.082N in the holding phase, despite the inevitable vibrations. The sensor has a low logical latency of 10ms which is suitable for real-time grasping applications.

The main advantage of the proposed approach is the combination of convolutional networks with recurrent layers which enables the sensor to estimate the contact force based on the object size relatively. The ConvLSTM architecture learns object geometry in early frames and estimate the contact force during a grasp.

Chapter 5

Data Augmentation

5.1 Introduction

Deep learning exhibits state-of-the-art performance in many fields including robotic grasping where a big data is required in training. However, deep learning models trained with small datasets commonly show worse performance. When large datasets are unattainable and expensive to generate, data augmentation can be a reasonable choice. In robotic grasping, small experimental datasets are sometimes common, and the problems to be solved have fewer input variables than that in other image recognition applications. Thus, deep learning is suitable for robotic grasping problems, such as tactile sensor for force estimation, and data augmentation is an effective and necessary method to achieve a generalised solution.

In chapters 3 and 4, a neuromorphic tactile sensor was presented to estimate the contact force using machine learning techniques. The proposed models were trained on all objects that are used in a grasp. Therefore, the experiments were conducted to collect data for each object for both training and validation purposes. Since performing experiments is required for each object size, the data collection procedure is time-consuming and costly for real-world applications.

To solve this challenge, augmentation techniques are proposed to generate artificial samples by perturbing existing samples in the training set. In this chapter, it is demonstrated that the augmentation methods improve the networks' accuracy without performing further experiments This approach reduces the cost and time of data collection process by creating a synthetic dataset from the conducted experiments. A new set of virtual experiments are performed and both image-based and temporal (time-domain) methods such as rotation, resize, and time shifting are implemented. A novel technique is proposed that shifts events across time dimension to generate

further synthetic samples. To evaluate each method, a network is trained on the original samples and the results are compared against the networks that trained on original and synthetic samples. The main contributions of this chapters are: (i) Development of time-domain and image-based augmentation for the neuromorphic tactile sensor to perform on objects with a different size. (ii) Proposing a novel event-based augmentation technique, "Temporal Event Shifting", to improve the sensor performance.

This chapter is organised as follows. The state-of-the-art studies of augmentation techniques are reviewed in section 5.2. The proposed image-based and time-domain augmentation methods are described in section 5.3. The experimental setup and procedures are presented in section 5.4. The results are demonstrated and discussed in section 5.5. Finally, the outcomes of this chapter are concluded in section 5.6.

5.2 Related Work

Data augmentation techniques aim to generate synthetic data for training of machine learning models. Specifically, in the supervised learning, the augmentation methods preserve the groundtruth for the generated samples. Augmentation techniques are divided into two main categories [151]: (i) Network-based augmentation; (ii) Algorithmic data manipulation. Network-based augmentation methods focus on creating networks to generate artificial samples from the real data such as Generative Adversarial Networks (GAN) introduced in [85].

Algorithmic data manipulation techniques apply fundamental operations on the data to generate realistic samples. For images, geometric transformation of the training data such as rotation, translation and shear has shown an improvement for classification tasks [152]. Another study [153], geometric translations and dropout layers are utilised to improve traffic signs recognition. The results indicate that the validation accuracy was improved by more than 5% considering rotation, translation and shearing augmentation methods. In addition to spatial methods, other image-based augmentation techniques such as image distortion, morphological, and noise injection techniques have increased the networks' accuracy for image classification [154].

In the augmentation process, many variables are involved which can be tuned by considering the real scenarios to achieve a network with a higher accuracy. In [155], a new framework (AutoAugment) is proposed to augment training data automatically. The proposed approach considers both feature-space and data-space augmentation methods to generate synthetic data. For validation, each experiment is performed three times to account for random initialisation of weights in the training process. From another point of view, effectiveness of refining the

labels for augmentation is investigated in [156]. The author demonstrates that general augmentation methods like cropping results in inaccurate labels for specific classes. Therefore, rules and conditions must be applied in augmentation process by considering samples of each class independently.

Time-series augmentation methods consider time and frequency domain features to generate artificial samples. One of the common approach in time-domain is shifting inputs in regards to the groundtruth. In [157], signals are shifted randomly to make the model robust against the shifted signals. Moreover, authors considered a combination of pitch shifting in frequency domain and time warping to improve the accuracy of the model for classifying environmental sounds. Window slicing is another popular approach in time-series classification which considers a sub-sample of original signal during both training and testing process of the model [158].

GANs are a class of machine learning models that includes two networks jointly learned to synthesise realistic artificial samples. The first network (known as generative) learns to generate samples from latent feature space while the second network (discriminator) identifies the originality of the produced samples. Figure 5.1 presents the concept of GAN architecture with generative and discriminator networks.

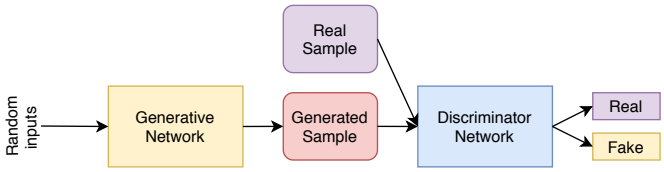


Figure 5.1: GAN structure with discriminator and generator networks to produce realistic artificial samples. The diagram is adopted from [159].

Although GAN achieved very interesting results in [85], there is a lack of stability for training in practice [160]. Several studies have modified the GAN structure to improve the generated samples. For instance, a cascade CNN with pyramid (multi-scale) features is proposed in [161] which has produced high-quality realistic samples. In [162], a novel class of architecture, Deep Convolutional Generative Adversarial Networks (DCGAN), is presented to generate samples in a unsupervised manner.

In addition to the image generation, time-dependent GANs are designed to capture temporal features and produce time-series samples. In [163], recurrent neural networks are employed in both generator and discriminator to produce continuous time-series samples. Similarly, recurrent conditional GAN is proposed in [164, 165] with conditions in time dimension to generate multi-dimensional time-series samples. A lot of time-series GAN with various architectures are designed recently, which are reviewed in [166] comprehensively.

GANs have shown a significant progress in computer vision applications recently. However, training a GAN requires a considerable number of training data to achieve acceptable results. Furthermore, training a GAN is a time-consuming process and often results are required to be confirmed by human. In algorithmic augmentation methods, features are tailored for the specific application based on logic that rules out impossible scenarios in the process.

Evaluation of the augmentation methods often is performed on validation set by using the augmented data in the training process. Since deep neural networks can easily overfit to the training data, validation performance provide more intuitive evaluation metric. For instance, algorithmic and GAN augmentation methods are used in [167] to evaluate the effectiveness of each method for a classification task on the validation set. Similarly, various augmentation methods are proposed in [168] to classify medical images. The the networks' accuracy are evaluated on the validation set to analyse the effectiveness of augmentation techniques.

Even though image augmentation techniques have been studied widely in the literature, no studies have been conducted to investigate event-based augmentation for different applications, including the tactile sensing applications presented in this thesis. It should be noted that the proposed augmentation techniques are completely different from event-based simulators such in [169] which aim to simulate events from the recordings of conventional cameras. In this chapter, time-domain and image-based augmentation methods are presented and compared to generate more data for a grasping task.

In the following section, image-based and time-series augmentation techniques are proposed to improve the network accuracy by generating synthetic samples.

5.3 Event Frame Sequence Augmentation

Events are characterised by location (x,y), timestamp and polarity. Similar to section 4.3.1, event frames are constructed by accumulation of events over a time window while preserving the spatial information. The sensor has a dimension of 240×180 which covers the contact area and the background. To reduce the memory requirements of the system and effect of the background noise, each frame is cropped to 140×150 pixels by considering the largest contact area size. Afterwards, the frames are downscaled to half (70×75) by adding closest neighbourhoods to a single pixel which results in a further reduction in the frame size. The resizing operation is performed by using OpenCV library on individual channels. Then, two channels are combined into one matrix to create the event frames. For visualisation purpose, the image is populated with the created matrix considering red and green channels. Figure 5.2 presents the cropping

and resizing process over the two channels.

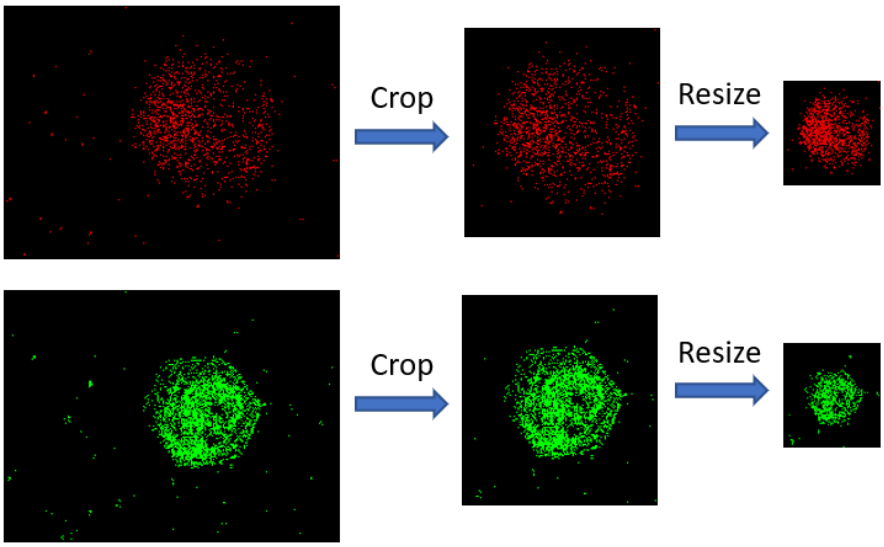


Figure 5.2: : Frames are constructed by accumulation events considering two channels for positive and negative polarities. Left images show the constructed frames while the middle and right images illustrate the constructed frames after cropping and resizing respectively.

After construction of the frames, the augmentation methods are applied to generate further artificial sequences for training the networks.

5.3.1 Image-Based Augmentation

Parallel grippers apply force on the object from both sides simultaneously, as shown in Figure 5.4. Therefore, the object orientation remains the same through the grasp after the object stabilization. Assuming that objects have the same shape, two main features are varied between different objects: (i) Size; (ii) Contact area orientation. Both of the feature variations can be implemented by affine transformations in the training data.

Rotation: Each grasping experiment, the contact area may be rotated around the central contact point. However, the object orientation remains the same through a grasp using parallel gripper. Hence, the same rotation transformation is applied on all frames rather than varying along the sequence. $X_t(x, y, p)$ represents the sequence of the original frames with spatial coordinates (x, y) with polarity p at timeframe t. For each experiment, the newly generated frames $X'_t(x', y', p)$ are formulated according to Equation 5.1.

$$X'_t(x', y', p) = X_t(x, y, p) \tag{5.1}$$

While Equation 5.2 represents the rotation around the centre of the object (x_o, y_o) by an angle

ϕ.

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \begin{bmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{bmatrix} \times \begin{bmatrix} x - x_o \\ y - y_o \end{bmatrix} \tag{5.2}$$

Resize: In order to augment the images to the desired size, the original images are required to be resized considering an specific scaling ratio β. The scaling ratio is determined based on the real object sizes where $\beta > 1$ and $\beta < 1$ for the resizing to the larger and smaller sizes respectively. We choose linear interpolation to assign values to the pixels. The resizing implementation is based on the OpenCV library, explained in [170]. Finally, a margin with zeros is considered to maintain the image size.

5.3.2 Noise

Noise in event-based applications is mainly derived from environmental light. To artificially apply noise on the training set, a set of experiments are recorded without any movements in the scene. Afterwards, the triggered events are considered as noise which are accumulated over a time window over two different channels. Finally, the frames in the training set are added to the noise frames to generate artificial samples.

5.3.3 Time-series Augmentation

In the grasping process, a lot of parameters such as DVS threshold, silicone material, sensor hysteresis and uncertainty cause a variable delay between the applied force and the triggered events. Time-series augmentation methods aim to generate artificial samples by considering transformations along the time dimension.

Frame Shifting: One of the simplest augmentation techniques in time domain is to shift frames index by a certain value (j) while preserving the groundtruth. This approach assists the network to deal with a slight lag between different experiments. Since shifting frames removes j frames from the input, new frames are required to be added to keep the sequence length fixed and are all set to zero values. Equation 5.3 presents the frame shifting process where the new frames are denoted as X'_t and j presents the shifting value. The frame shifting is applicable in two directions (i) Left: The frames are shifted to the earlier timestamps ($j < 0$); (ii) Right: The frames are shifted to the future timestamps ($j > 0$).

$$\forall t, \quad X'_t(x, y, p) = X_{t+j}(x, y, p) \tag{5.3}$$

Temporal Event Shifting: Similar to the frame shifting, we propose a novel approach to shift events across the frames, called "Temporal Event Shift (TES)". The proposed method selects a fraction ζ of events ($0 < \zeta < 1$) randomly in each frame. These events are removed from the current frame and added to the next or previous j frames. Figure 5.3 demonstrates the procedure for temporal event shifting to right while preserving the spatial information of events.

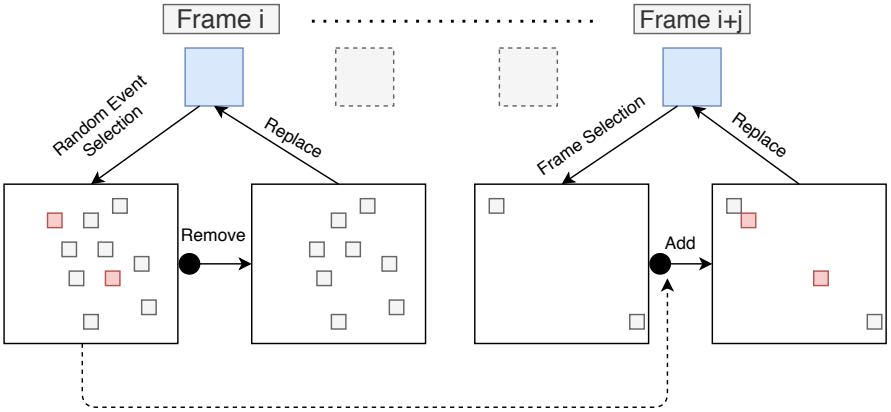


Figure 5.3: : Temporal event shifting diagram when a ratio of events is shifted to the future frames (j>0).

To shift the events to the past frames, j value is considered as negative number. This process is formulated in Equation 5.3 where the new frame is denoted as X'_t. $\forall t, p$, create randomly a frame $Z_t(x, y, p)$ such as:

$$Z_t(x, y, p) \leq X_t(x, y, p), \quad \forall x, y \tag{5.4}$$

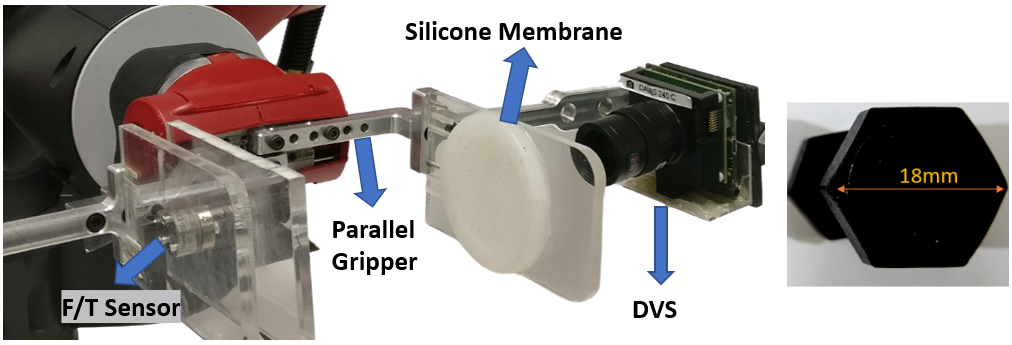
$$\sum_{x,y} Z_t(x, y, p) = \zeta \cdot \sum_{x,y} X_t(x, y, p) \tag{5.5}$$

$$X'_t(x, y, p) = X_t(x, y, p) - Z_t(x, y, p) + Z_{t+j}(x, y, p) \tag{5.6}$$

5.4 Experimental Setup

Similar to chapter 4, the real experiments are conducted on a Baxter robot including F/T sensor, silicone membrane, DVS, and 3D printed transparent planes. The transparent silicone membrane has 50 shore hardness and 8mm depth. Furthermore, the range of contact force is set to 0-25N which is significantly higher than the force range in chapter 3-4. Figure 5.4(a) presents the experimental setup for the grasping task.

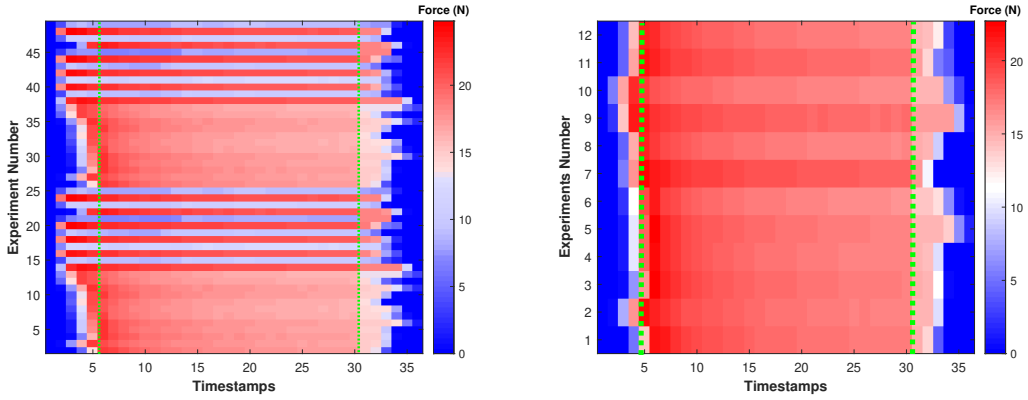
Similar to chapters 3-4, three bolts with 12, 15 and 18mm diameter are used for the grasping process as show in Figure 5.4(b). However, the time for each phase is different from the other



(a) (b)

Figure 5.4: :(a) A DVS is mounted on the left plane to observe the intensity changes in the contact area through the silicone membrane. A F/T sensor is located on the right plane to record force values through the grasp. (b) A bolt with 18mm diameter painted in black.

chapters. For splitting the data, we choose the small and large bolts for the training (48 sequences) while the medium bolt experiments (12 sequences) are considered for the validation. Figure 5.5 presents the force values recorded by F/T sensor for the training (a) and validation sets (b).



(a) (b)

Figure 5.5: Each row demonstrates the force values that is captured by the F/T sensor over time. (a) Training set: 48 experiments are conducted using the small and large objects. (b) Validation set: 12 experiments are considered considering the medium size object.

Two configurations are set for the gripper to grasp the object with a different applied force. The experimental setup is not fully controlled which results in a slight variation of force between experiments with the same configuration. Therefore, a slight variation of force over time is visible.

5.4.1 Preparation of Frames

The experiments have the maximum length of 360ms. In this chapter, 36 frames are conducted for each experiment by accumulation of events over a 10ms window. The frames are cropped to 140x150 to reduce the noise and eliminate the background which is selected based on the largest object contact area. Afterwards, the frames are resized to 70x75 pixels considering the accumulation of neighborhoods. The resizing ratio is selected based on the maximum saturation level of each pixel over the time window. The force readings have a resolution of 2ms which is measured by the F/T sensor. After the synchronization, force measurements are read every 10ms to synchronize them with the frames.

5.4.2 Training Configurations

In chapter 4, a variety of LSTM, CNNLSTM and ConvLSTM architectures are tuned to find the most accurate network. The ConvLSTM architecture with four ConvLSTM hidden layers have achieved the least error for the force estimation. In this section, we choose the same architecture to compare effectiveness of the augmentation techniques on the networks' accuracy.

Adam optimizer is used to minimize the training loss (MSE) for the training set while monitoring the validation loss for selecting the best network. The training process finishes when the validation loss stops improving after 20 consecutive epochs. All the models are trained with the same configuration to provide a fair comparison. Keras framework is used to set the training configuration using an NVIDIA 1080 GPU.

As demonstrated in Figure 5.5, the experiments are divided to two sets: training (48 sequences) and validation (6 sequences). The augmentation methods increase the volume and variation of training set to improve the accuracy of the trained model. To evaluate the effectiveness of each augmentation technique, the training set is doubled (96 sequences) by synthesising as many samples as the real ones.

5.5 Results and Discussion

To evaluate the augmentation techniques, the training data size is doubled with the artificial samples while preserving the groundtruth. Since the random initialisation of weights affect the training process, the random seed is controlled for 10 runs. The final results is obtained by averaging the errors over the lowest error on the validation set using the same random

initialisation. Figure 5.6 presents the average of MSE for the validation set where the red line shows the standard deviation of MSE for the image-based augmentation methods.

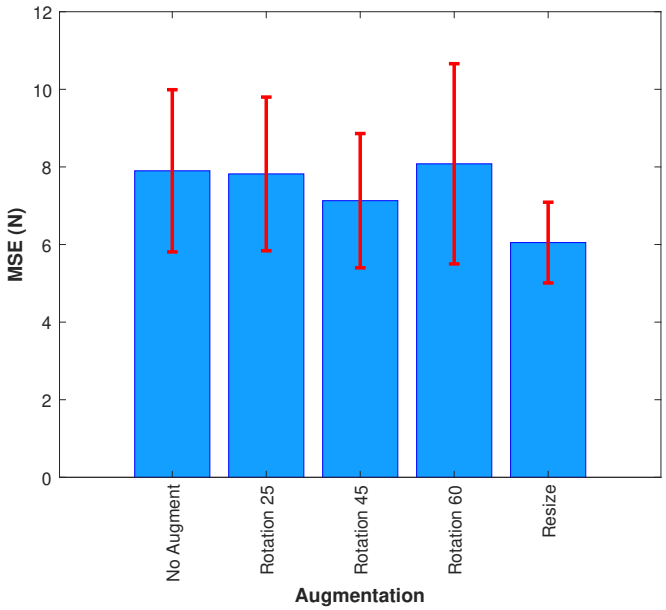


Figure 5.6: : MSE for geometric augmentation methods. y-axis shows the average of MSE(N) for the trained networks after 10 repetition with random initialisation. The red lines represent the standard deviation of MSE(N) for each method.

The network trained without augmentation (No Augment) achieves the MSE of 7.89N with STD of 2.09N. The standard deviation of more than 25% indicates instability of training process with respect to the random initialisation. The rotation of images between 0 and 45 degrees (Rot45) provides a slight improvement in network accuracy. The best result from the geometric augmentation approaches is achieved by the resizing method for the desired object size. The scaling factor of resizing is considered as 1.25 and 0.83 for small and large objects respectively.

On the other hand, we consider a background noise for further augmentation. The background noise includes both events polarities which are added the original frames to double the training samples. The results indicate an slight improvement of 10% in MSE and standard deviation compared to the networks that are trained without augmentation.

The results indicate that the MSE of network is reduced to 6.05N and the standard deviation is decreased to 1.04N, a decrease of 50%. Therefore, the resizing augmentation method is the most effective image-based augmentation method, which makes sense as the challenge in our experiments was to train the networks for an unseen object size.

Two time-series augmentation methods, mentioned in section 5.3.3, are tested: Frame Shifting (FS) and the proposed Temporal Event Shifting (TES). For the FS method, j is varied between -3 and 3 to find the most effective value to shift the frames. Figure 5.7 presents the effectiveness

of frame shifting augmentation with different j values.

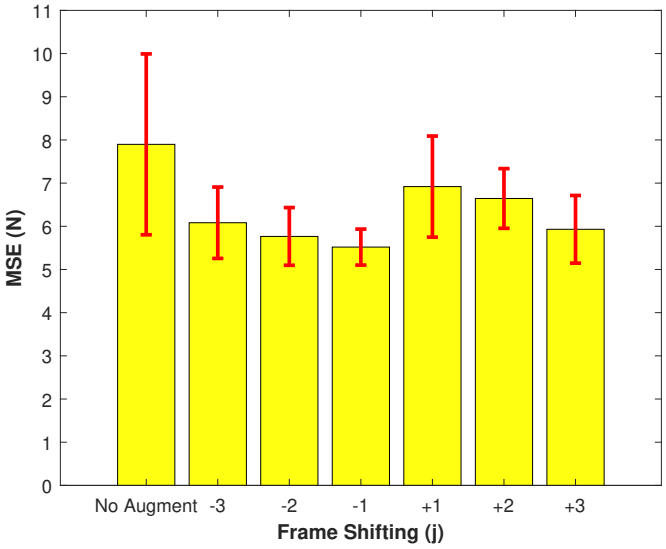


Figure 5.7: Comparison of average MSE for frame shifting methods. x-axis shows the j value for frame shifting and the red bar illustrates the standard deviation of MSE over 10 runs.

Shifting one frame to left (FS-1) results in the lowest MSE of 5.51N which is 30% less than the MSE achieved without augmentation. Furthermore, the STD of errors has reduced significantly to one fourth (0.41N) of the networks trained on the real data.

For the TES method, the ratio of the events selection (ζ) is considered as 0.25, 0.50 and 0.75 with the same j variations as in the FS method. Figure 5.8 demonstrates the average MSE of the validation set considering different j and ζ. Among the TES augmentation configurations, two frames shift to the left with 50% threshold (TES-2(0.50)) results to the minimum MSE of 5.98N with 30% reduction of standard deviation (0.53N) compared to the results without augmentation.

In FS-based augmentations, the amount of new data generated is limited to one new sample for original sample considering a fixed j value. On the other hand, in TES-based augmentations, the random seeds affect the selection of events, and as a consequence an unlimited number of new samples can be produced for specific values of j and ζ. We produced an experiment to generate 480 artificial samples by varying the seed for FS-2(50) method. The results show that increasing the generated samples does not improve the network performance where an average MSE of 6.25N with 0.82N standard deviation is achieved. The main reason for this phenomenon is that the groundtruth remains the same, despite the significant variation in the input.

Most of the augmentation techniques in time domain improve the networks' performance. The main factors that affect the events through the time dimension are the F/T sensor hysteresis,

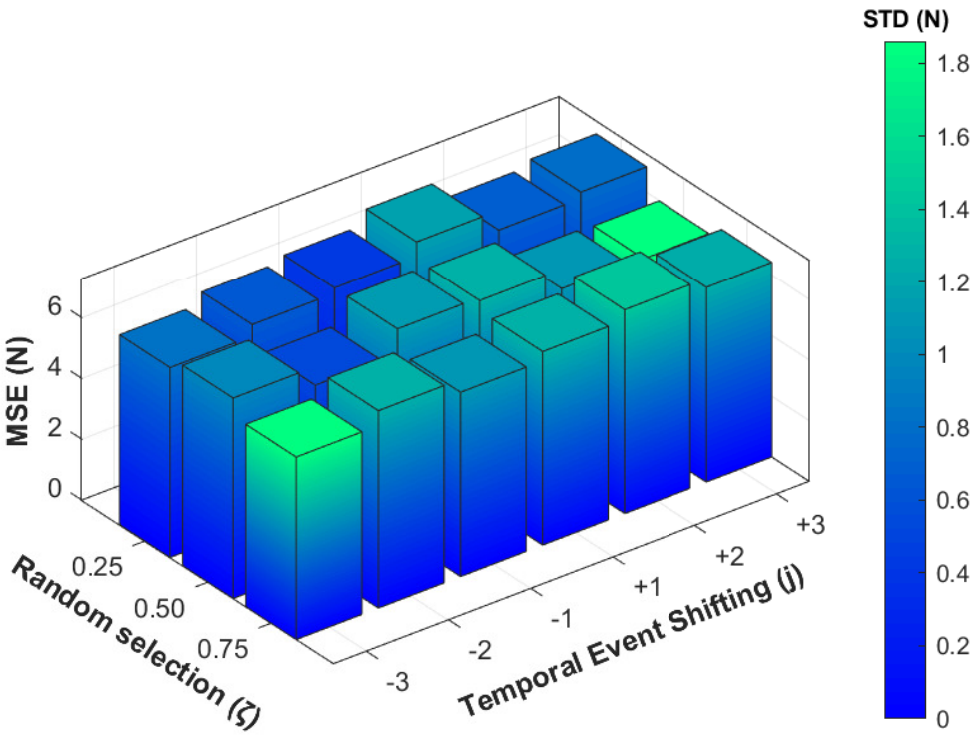


Figure 5.8: Comparison of average MSE of the networks for Temporal Event Shifting augmentation technique. x-axis and y-axis present the j and ζ values respectively. The MSE value of each method is illustrated on z-axis. The color of each bar surface represents the standard deviation (STD) of MSE over 10 runs.

non-linear behaviour of the silicone membrane, uncertainty and vibrations. These factors are inevitable in a real-world applications which shows the benefit of the augmentation methods in the time domain.

Due to the increase of the sensor range from 3N to 25N, the results of this chapter have a considerably higher error compared to chapter 3-4 results. The main factors for this high error are the elastomer elasticity and DVS threshold. Design of a new elastomer by targeting the contact force range will improve the sensor performance significantly.

As discussed in chapters 3 and 4, a grasp is divided into three phases. The grasping phase is defined where the contact force increases to the maximum level (The first 5 frames). The holding phase includes a slight variation of force during the time from 6th frame to 30th frame. Finally, the releasing phase where the force values are decreased continuously to zero (The last 5 frames). Figure 5.9 presents the average of estimated force (blue) and groundtruth (red) for two examples of the validation set over 10 runs. The top row (a,b) demonstrates the average of the force predictions for training without augmentation while the middle row (c,d) presents the the average of estimated force considering FS-1 method. The bottom row (e,f) demonstrates the

average of estimated force and groundtruth using TES-2(0.50) method. The highlighted area illustrates for the standard deviation of the estimated force over 10 runs. The phases of a grasp are differentiated by the green line in each figure.

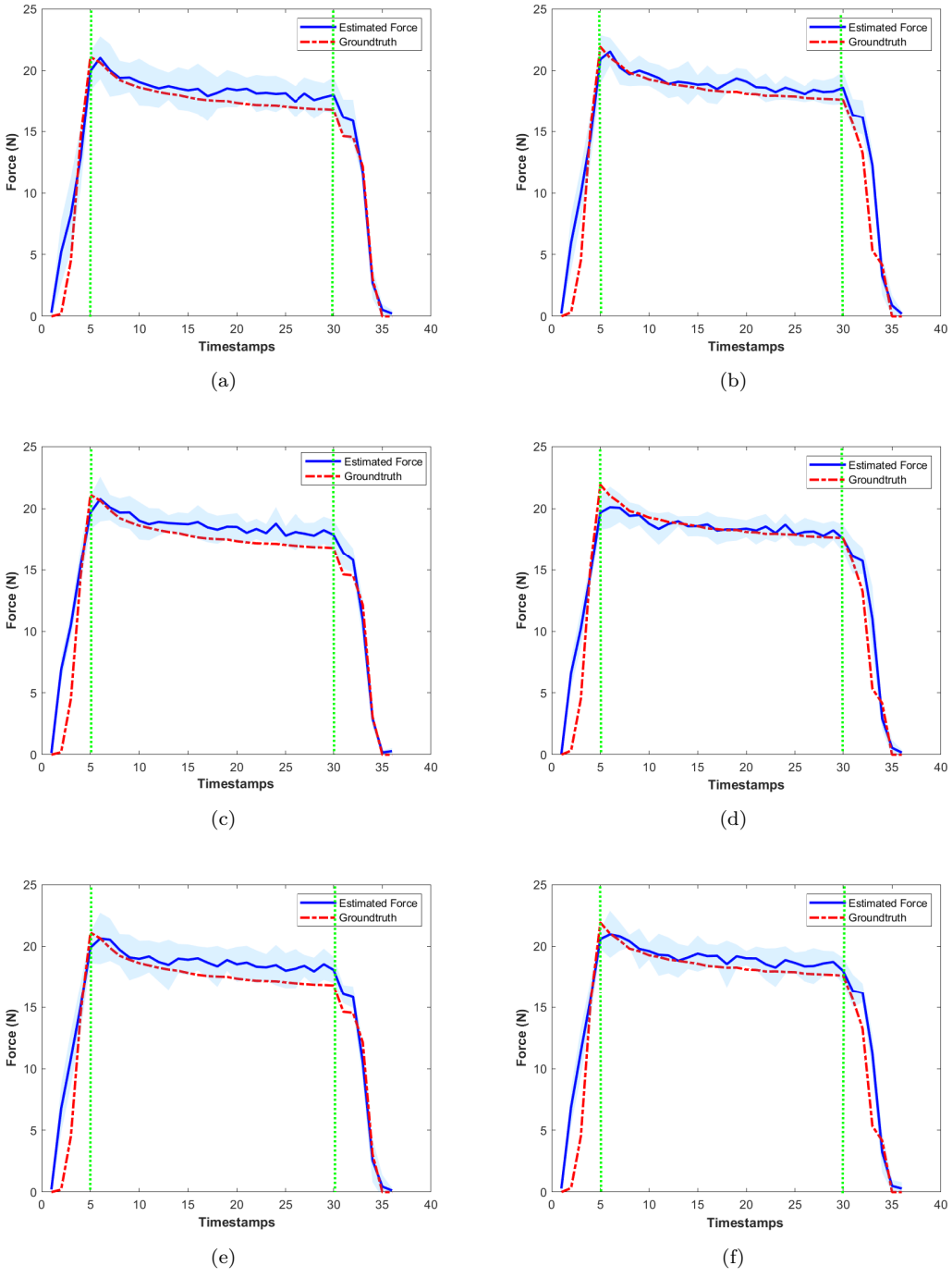


Figure 5.9: Each column presents an experiment from the validation set.Top row presents the average of estimated force and grountruth without augmentation. The middle row demonstrates the output of the network for FS-1 augmentation method. The bottom row (e,f) presents the average of estimated force and groundtruth for TES-2(0.50) augmentation method.

The results indicate that both frame shifting and temporal event shifting augmentation reduce the standard deviation of the predictions in all three phases. In fact, the impact of random

initialisation is decreased by augmenting the training data. In Figure 5.9(b) and (d), a clear improvement of the estimated force in the most part of the vibration phase is visible. Even though the frame shifting results in a lower MSE and standard deviation, temporal event shifting method captures the maximum contact force (at 5^{th} timestamp) more accurately in most of the cases.

In order to investigate the impact of augmentation methods on all the measurements, 12 predictions of 10 models are considered for grasping, holding and releasing phases. The final results include 4320 points which are demonstrated in Figure 5.10. The black line presents the contact force measured by F/T sensor. The estimated force are presented by cross while blue, red and green are for grasping, holding, and releasing phases respectively. Figure 5.10 compares the estimated force using FS-1 and TES-2(50) augmentation techniques.

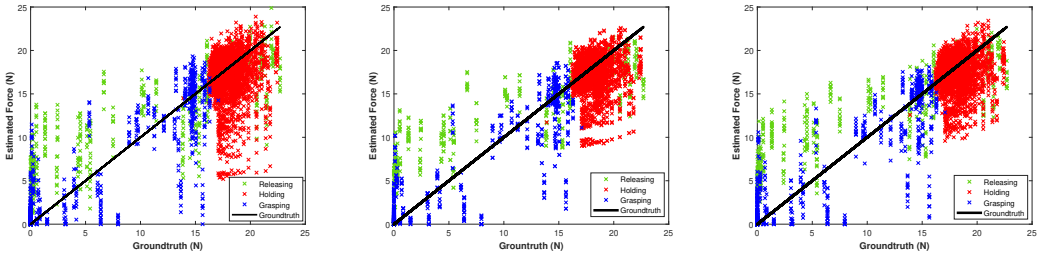


Figure 5.10: (a) shows the estimated force for the model without augmentation. Figure (b) and (c) demonstrates the estimated force for the model which is trained with FS-1 and TES-2(50) augmentation techniques respectively.

As observed in Figure 5.10, FS-1 and TES-2(50) augmentations improve the force estimation in the holding phase. Both augmentation methods shift events to the earlier frames to create artificial samples. The main reason of this phenomenon is that the number of triggered events increases significantly after applying the certain amount of force. Therefore, shifting the events to the left allow the network to relate more events to the contact force in the early frames. Furthermore, the silicone membrane has a non-linear deformation which absorbs a ratio of the contact force particularly in the transition phases. The force absorption coupled with the F/T sensor hysteresis introduce a variable delay between the triggered events and the contact force.

The image-based and time-domain augmentation methods synthesise the training data from different prospective. Therefore, a combination of both methods provide both spatial and time-domain feature in the generated samples. Since the best accuracy is achieved by resizing and FS-1, these two methods are combined to generate a new set of synthetic samples. There are two ways to combine the two methods: (i) Perform each augmentation method independently to generate artificial samples ;(ii) Hybridise both augmentations methods on samples to generate

a set of synthetic samples. The results indicate that independent augmentation of each sample achieve a better accuracy than simultaneous combination of methods. The independent sample generation method reduces the average MSE of the networks to 5.71N with standard deviation of 1.06N which is slightly higher than FS-1 method. The hybrid augmentation method results in a high MSE of 7.20N with standard deviation of 1.22N, significantly higher error compared to FS-1 method. Figure 5.11 demonstrates the average MSE of the proposed augmentation methods where the standard deviation is highlighted as a red line.

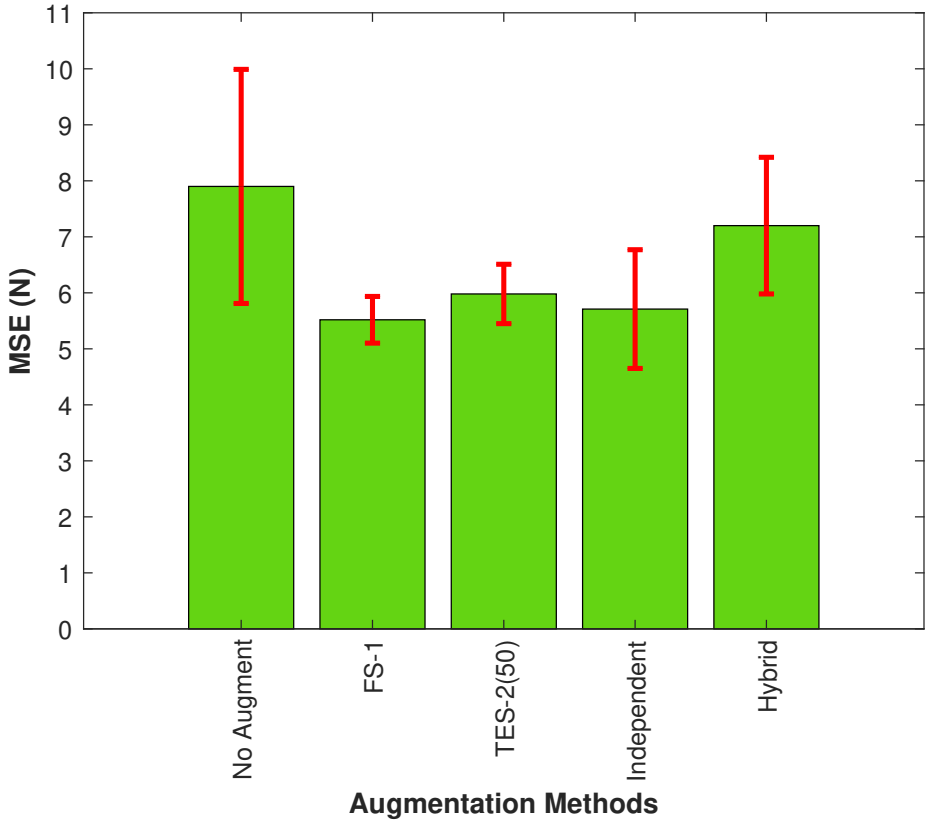


Figure 5.11: Comparison of average MSE for the proposed augmentation method. The independent and hybrid methods are considered for FS-1 and resizing methods that achieved the lowest error for time-domain and image-based techniques respectively. The standard deviation of each method is presented as a red line.

In the image-based augmentation techniques, resizing the object to a desired size results in the best accuracy. Since the network learns the relationship between the applied force and triggered events based on the contact area, resizing the training data simulates the experiments for the new size of an object.

In chapter 3, a noticeable delay was observed in the releasing phase where the network always respond faster than the F/T sensor. Similarly, In chapter 4, a high standard deviation and

error are demonstrated on the releasing phase. In fact, the elasticity of silicone membrane has a significant impact on the delay between the triggered events and the contact force. Therefore, the augmentation methods in time-domain improve the network accuracy remarkably whereas FS-1 results in the lowest average of MSE.

5.6 Conclusion

In this chapter, augmentation methods are investigated to improve the force estimation using an object with a different size. Image-based and time-domain augmentation techniques are implemented to add new training samples by perturbating existing ones, aiming to improve the network accuracy. The main contributions of this chapters are: (i) Development of both spatial and temporal augmentation techniques to increase the sensor accuracy for the objects with a different size. (ii) Proposing a novel event-based augmentation technique, "Temporal Event Shifting", to synthesise data temporally while preserving the spatial information of the events.

The results indicate that both time-domain and image-based augmentations improve the network accuracy. The best results are achieved by synthesising training data with FS-1 which improves the network accuracy by approximately 30%. This is followed by independent augmentation using FS-1 and resizing methods with MSE of 5.98N, almost similar performance to FS-1. Therefore, the proposed methods reduce the number of required experiments which decrease the cost and time of data collection process.

Chapter 6

Conclusion

6.1 Introduction

In summary, this thesis proposed a novel neuromorphic vision-based tactile sensor for the contact force measurements and material classification. This novel sensor captured the intensity changes within the contact area between objects and the gripper soft membrane. The intensity changes were modelled into the contact force measurements and material classification using machine learning techniques. This thesis is the first work that demonstrates the utilisation of neuromorphic vision sensors in tactile sensing for the contact force measurements and material classification.

A comprehensive review of tactile sensors, neuromorphic vision sensors and relevant machine learning techniques was provided in chapter 2. This is followed by proposing a novel neuromorphic vision sensor in chapter 3 to estimate contact force and classify materials using Time-Delay Neural Networks (TDNN) and Gaussian Process (GP). Afterwards, spatio-temporal deep learning models such as Convolutional LSTM (ConvLSTM) were investigated in chapter 4. The spatio-temporal networks achieved a higher accuracy compared to temporal networks. Finally, a novel technique was proposed in chapter 5 considering both spatial and temporal features of the events to augment the dataset. Use of data augmentation improved the networks' accuracy without the need to perform further real experiments.

The review of the state-of-the-art solutions that tackle this problem has brought up the following research questions:

1. How a neuromorphic vision sensor can be used to acquire tactile information?

2. What are the suitable machine learning techniques for the contact force estimation and material classification?

3. How synthetic data can be generated to improve machine learning accuracy for the tactile measurement systems?

These questions were thoroughly investigated to enhance vision-based tactile sensor which is the aim of this thesis. Accordingly, the contributions of this thesis have been made for question 1 in section 6.2.1, question 2 in section 6.2.2-6.2.3, and the third question in section 6.2.4.

The rest of this chapter is structured as follows: at first, the contributions stated in chapter 1 are revisited in section 6.2. This is followed by proposing directions for future work in section 6.3 and epilogue in section 6.4 respectively.

6.2 Summary of Contributions

This thesis presented and validated the following novel contributions:

1. The first neuromorphic vision-based sensor to measure the contact force and classify materials in a grasp using a neuromorphic camera (DVS).

2. A Time Delay Neural Network (TDNN) and a Gaussian Process (GP) to find the correlation between the triggered events and the contact force.

3. A deep Long Short-Term Memory (LSTM) network with convolutional layers to estimate the contact force from spatio-temporal event data.

4. Geometric and time-domain augmentation techniques applied on spatio-temporal event data to enhance the force estimation accuracy for an unseen object.

The following sections discuss these contributions in more details.

6.2.1 Neuromorphic Vision-based Tactile Sensor

A thorough review of literature for vision-based tactile sensors was presented in section 2.1.2.1 and advantages and disadvantages of each method were highlighted. This is followed by more narrower review of vision-based methods in sections 3.2 and 4.2 which indicates that this thesis

is the first work to develop a neuromorphic vision-based sensor for contact force estimation and material classification.

In chapter 3, a general framework was proposed to demonstrate the elements of a neuromorphic vision-based tactile sensor. This is followed by a description of sensor operations in 3.4.1 which indicates a relationship between the contact force and triggered events. This relationship was investigated for a robotic grasping task in 3.4.2 which shows the correlation of positive and negative events with the contact force variations. The experiments in chapter 3-5 validated the relationship between the contact force and events by applying different machine learning methods to correlate the triggered events with the contact force.

The main advantages of the proposed sensor are high temporal resolution and wide dynamic range compared to other vision-based tactile sensors. The high temporal resolution enables the sensor to capture intensity changes with a low latency which is necessary for real-time applications. The wide dynamic range increases the sensor sensitivity to perform in challenging environments with low-light conditions. Furthermore, neuromorphic vision sensors offer a low power consumption and memory requirements which are significantly important parameters for robotic applications.

6.2.2 Temporal Force Estimation and Material Classification

Chapter 3 also presented a novel method to estimate the contact force and classify materials from triggered events. Initially, the number of events are accumulated over time to formulate the input of the system, which also provides a memory. Afterwards, TDNN and GP models were applied to estimate the contact force. In addition, a DNN was implemented to classify materials in a grasp. Forty-eight experiments were performed considering Foam, Silicone, Rubber and Steel materials with the same shape to validate the proposed methods. The modelling parameters for all three models were tuned and the results were cross-validated against the FlexiForce-A201 sensor using leave-one-out method.

The results indicate that the contact force can be estimated in grasping and releasing phases with high accuracy based on temporal event data. The logical delay of TDNN was 21ms (approximately 47FPS) which is lower than conventional cameras sampling rate. Moreover, object materials were identified by the proposed DNN after a single grasp. The classification accuracy of 79.17% was achieved with the neuromorphic sensor, which is almost 30% higher than the classification of piezoresistive measurements, using a similar methodology.

6.2.3 Spatio-temporal Force Estimation

In chapter 4, a novel technique was proposed to estimate the contact force using spatio-temporal event data. The framing algorithm was formulated in section 4.3.1 to preserve the spatio-temporal information of the events. Afterwards, three LSTM-based networks, namely, CNNLSTM, ConvLSTM and LSTM, were developed to estimate the contact force in a grasp.

The methods were tuned and cross-validated against a ATI Nano F/T sensor using leave-one-out method. The results showed that ConvLSTM outperformed CNNLSTM and LSTM networks. In fact, spatio-temporal event data assisted the network to relate the contact force to the events based on the size of the contact area. Furthermore, a deep analysis of results was provided in section 4.6 while the similarity measurements of the network estimations and groundtruth were compared.

This contribution demonstrated that in addition to temporal event data, spatial event data improves the modelling of the intensity changes into the contact force. Therefore, the networks with spatio-temporal modelling capabilities such as ConvLSTM and CNNLSTM learn the relationship between the triggered events and the contact force robustly.

6.2.4 Data Augmentation

The proposed neuromorphic vision-based tactile sensor in chapter 3 and 4 was developed based on deep learning methods to estimate the contact force and classify materials. To reduce the cost and time of data collection process, augmentation techniques were investigated in chapter 5. Assuming change of the object size in the experiments, time-domain and spatial-domain methods were proposed to augment the training data. In spatial domain, rotation and resizing techniques were investigated where the resizing method improved the network accuracy significantly. On the other hand, Frame Shifting (FS) technique was implemented to synthesise the training data in time-domain. Furthermore, a novel approach, Temporal Event Shifting (TES), was proposed to augment events across the time-domain while preserving the spatial information.

Sixty experiments were performed by considering three objects with the same shape, but different sizes (small, medium and large). The training data consisted of large and small objects while the medium object was selected for the validation set only. A ConvLSTM network was trained 10 times with random initialization using both real and augmented data to investigate the impact of augmentation. The results indicated that FS-1 and combination of TES-2(50) with resizing method achieve a competitive outcome by improving the network accuracy by

approximately 30%.

For a practical application, the sensor must be capable of accurate measurements for unseen objects. Hence, a large data set is required to be collected for objects with various sizes and shapes. The proposed augmentation methods reduce the required number of experiments for the sensor. For example, experiments for 10 different sizes of a specific object will be reduced to two sets of experiments only. The experiments for the largest and smallest objects may be used to augment data for the other 8 different sizes considering the proposed augmentation techniques. Therefore, time and cost of data collection will be decreased significantly.

6.3 Limitations and Future Work

This thesis proposed the first neuromorphic vision-based tactile sensor for the contact force estimation and material classification. Thus, several future research directions are suggested in this relatively new field as follows.

Membrane Design and Calibration: As reviewed in chapter 2, design and material of the sensor membrane impacts significantly the sensor performance. A wide variety of membranes may be designed for different applications in regard to range of force, size of gripper, object shape and sensor sensitivity. Since neuromorphic vision sensors fire events based on thresholding the intensity changes, a benchmark can be conducted to optimise the events threshold by considering an specific range of force. This approach will lead to create a calibration process to increase traceability of the measurements for different membranes.

Event Representation: In chapter 3, the events were accumulated over time for each experiment to be used for machine learning methods to estimate the contact force. This approach is limited to relatively short sequences since longer sequences with a lot of force variations will saturate the accumulation of events curve. The proposed methods are based on the accumulation of events without restarting the state of the system. For a long sequence, derivative of events accumulation becomes insignificant after certain period. Therefore, the learning algorithm cannot converge to the optimal point due to the vanishing gradient problem. Consequently, the predicted results would not be accurate for a long sequence. To address this problem, a long sequence can be broken down to multiple sequences with a state for each part of the signal. The state of the sensor imitates the memory for the sensor to link the current sequence with a final force value of the previous sequence. This procedure can be implemented in real time after prediction of the contact force for each sequence. In addition, other techniques such as weighted

accumulation of events and time surfaces [171] may be developed to deal with longer sequences or continuous online operation.

Machine Learning: Spatio-temporal LSTM-based networks proposed in chapter 4 has demonstrated improvement in the networks which are using memory and spatial features. However, the experiments were performed with a similar length by repeating a grasp with the same configurations. This approach limits the sensor functionality for handling sequences with multiple grasps. For future work, a sliding window approach can be considered to train a network on temporal windows for longer sequences with stateful configuration.

On the other hand, attention-based deep learning models have become popular for learning long sequences in computer vision and natural language processing [172]. Therefore, attention-based models may be investigated for the contact force estimation as well as texture or material classification. As reviewed in section 2.2, in the paradigm of neuromorphic signal processing, algorithms are divided into processing signals event-by-event or events as groups. This thesis considered the latter approach of processing events as a group by constructing frames over time. However, Spiking Neural Networks (SNN) may be implemented to process event-by-event to achieve an end-to-end neuromorphic system. Finally, it is noteworthy to mention that for future work, the proposed networks may be trained in a multi-task manner to reduce number of models into one general model for the contact force estimation, material and texture classification.

3D Force Measurements: In thesis, the force measurements were estimated along one dimension (normal force) for parallel grasping. However, three-dimensional force measurements provide further information about the contact area in order to feedback controllers for correction of grasp. The same principle of the proposed sensor in this thesis may be extended by performing experiments on humanoid grippers to apply force in different directions. Since the membrane deforms in regard to the contact force vectors, the events will be triggered with a different pattern. Therefore, a supervised machine learning method can be implemented to estimate the contact force in three dimensions.

Data Augmentation: As shown in chapter 5, augmentation of event-frames improved the network accuracy significantly. However, the groundtruth remained the same for all experiments which limits the generalisation of augmentation techniques. A novel approach may simulate the contact area using modelling software for deformable materials. Afterwards, the relationship between the contact force and events may be acquired from the real experiments. Finally, the simulations may be represented by events which can be used for training purposes, similar to the proposed framework in [173]. Furthermore, a deep learning method such as Generative Adversarial Networks may be adapted to generate synthetic samples for material or texture

classification as well as contact force estimation.

In chapter 3-5, the objects are coloured in black to increase the contrast between the contact area and background. To generalise the sensor for objects with different colours, a black layer of rubber can be attached to the surface of the elastomer to eliminate the noise. Furthermore, the speed of the grasping has a significant impact on the triggered events in a fixed time-window. Therefore, further experiments with a variety of speed can be performed to improve generalisation of the sensor for different speeds.

6.4 Epilogue

This thesis successfully proposed a novel class of vision-based tactile sensors, the neuromorphic vision-based tactile sensor, to measure the contact force and classify materials in a grasp. The proposed sensor relied on modern machine learning methods to model intensity changes within the contact area into the contact force. This novel sensor was validated through different sets of experiments in the thesis. It was demonstrated that both temporal and spatial event data of the contact area can be used to model the soft membrane deformation, and therefore, the contact force measurements. The main advantages of the proposed sensor are high temporal resolution and wide dynamic range which enable the sensor to perform robustly in real-time applications. Moreover, novel augmentation techniques were proposed to reduce the time and cost of data collection process. The results were promising and this novel sensor could be adapted for practical robotic applications in future.

References

[1] R. S. Johansson and G. Westling, "Roles of glabrous skin receptors and sensorimotor memory in automatic control of precision grip when lifting rougher or more slippery objects," *Experimental Brain Research*, vol. 56, no. 3, pp. 550–564, 1984.

[2] R. S. Johansson and K. J. Cole, "Grasp stability during manipulative actions," *Canadian Journal of Physiology and Pharmacology*, vol. 72, no. 5, pp. 511–524, 1994.

[3] T. Mukai, M. Onishi, T. Odashima, S. Hirano, and Z. Luo, "Development of the tactile sensor system of a human-interactive robot "RI-MAN"," *IEEE Transactions on Robotics*, vol. 24, no. 2, pp. 505–512, 2008.

[4] Y. Chebotar, O. Kroemer, and J. Peters, "Learning robot tactile sensing for object manipulation," *IEEE International Conference on Intelligent Robots and Systems*, no. Iros, pp. 3368–3375, 2014.

[5] P. Puangmali, K. Althoefer, L. D. Seneviratne, D. Murphy, and P. Dasgupta, "State-of-the-art in force and tactile sensing for minimally invasive surgery," *IEEE Sensors Journal*, vol. 8, no. 4, pp. 371–380, 2008.

[6] P. Puangmali, H. Liu, L. D. Seneviratne, P. Dasgupta, and K. Althoefer, "Miniature 3-axis distal force sensor for minimally invasive surgical palpation," *IEEE/ASME Transactions on Mechatronics*, vol. 17, no. 4, pp. 646–656, 2012.

[7] M. R. Cutkosky, "On Grasp Choice, Grasp Models, and the Design of Hands for Manufacturing Tasks," *IEEE Transactions on Robotics and Automation*, vol. 5, no. 3, pp. 269–279, 1989.

[8] A. Bicchi and V. Kumar, "Robotic grasping and contact: A review," *Proceedings-IEEE International Conference on Robotics and Automation*, vol. 1, pp. 348–353, 2000.

[9] P. Sabetian, A. Feizollahi, F. Cheraghpour, and S. A. A. Moosavian, "A compound robotic hand with two under-actuated fingers and a continuous finger," *9th IEEE International Symposium on Safety, Security, and Rescue Robotics, SSRR 2011*, pp. 238–244, 2011.

[10] L. Wang, J. DelPreto, S. Bhattacharyya, J. Weisz, and P. K. Allen, "A highly-underactuated robotic hand with force and joint angle sensors," *IEEE International Conference on Intelligent Robots and Systems*, pp. 1380–1385, 2011.

[11] R. Deimel and O. Brock, "A novel type of compliant and underactuated robotic hand for dexterous grasping," *International Journal of Robotics Research*, vol. 35, no. 1-3, pp. 161–185, 2016.

[12] A. Cirillo, F. Ficuciello, C. Natale, S. Pirozzi, and L. Villani, "A conformable force/tactile skin for physical human-robot interaction," *IEEE Robotics and Automation Letters*, vol. 1, no. 1, pp. 41–48, 2016.

[13] H. Shen, "The soft touch," *Nature*, vol. 530, no. 7588, p. 24, 2016.

[14] H. Liu, D. Guo, and F. Sun, "Object Recognition Using Tactile Measurements: Kernel Sparse Coding Methods," *IEEE Transactions on Instrumentation and Measurement*, vol. 65, no. 3, pp. 656–665, 2016.

[15] S. Luo, J. Bimbo, R. Dahiya, and H. Liu, "Robotic tactile perception of object properties: A review," *arXiv*, vol. 48, no. August, pp. 54–67, 2017.

[16] V. Maheshwari and R. Saraf, "Tactile devices to sense touch on a par with a human finger," *Angewandte Chemie - International Edition*, vol. 47, no. 41, pp. 7808–7826, 2008.

[17] R. S. Dahiya, G. Metta, M. Valle, and G. Sandini, "Tactile sensing-from humans to humanoids," *IEEE Transactions on Robotics*, vol. 26, no. 1, pp. 1–20, 2010.

[18] Z. Kappassov, J. A. Corrales, and V. Perdereau, "Tactile sensing in dexterous robot hands - Review," *Robotics and Autonomous Systems*, vol. 74, pp. 195–220, 2015.

[19] H. B. Muhammad, C. Recchiuto, C. Oddo, L. Beccai, C. J. Anthony, M. J. Adams, M. C. Carrozza, and M. C. Ward, "A capacitive tactile sensor array for surface texture discrimination," *Microelectronic Engineering*, vol. 88, no. 8, pp. 1811–1813, 2011.

[20] J. G. V. da Rocha, P. F. A. da Rocha, and S. Lanceros-Mendez, "Capacitive sensor for three-axis force measurements and its readout electronics," *IEEE Transactions on Instrumentation and Measurement*, vol. 58, no. 8, pp. 2830–2836, 2009.

[21] S. Stassi, V. Cauda, G. Canavese, and C. F. Pirri, "Flexible tactile sensing based on piezoresistive composites: A review," *Sensors (Switzerland)*, vol. 14, no. 3, pp. 5296–5332, 2014.

[22] Y. Jung, D. G. Lee, J. Park, H. Ko, and H. Lim, "Piezoresistive tactile sensor discriminating multidirectional forces," *Sensors (Switzerland)*, vol. 15, no. 10, pp. 25 463–25 473, 2015.

[23] S. Teshigawara, K. Tadakuma, A. Ming, M. Ishikawa, and M. Shimojo, "High sensitivity initial slip sensor for dexterous grasp," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 4867–4872, 2010.

[24] P. Payeur, C. Pasca, A. M. Cretu, and E. M. Petriu, "Intelligent haptic sensor system for robotic manipulation," *IEEE Transactions on Instrumentation and Measurement*, vol. 54, no. 4, pp. 1583–1592, 2005.

[25] A. Sanli, R. Ramalingame, and O. Kanoun, "Piezoresistive pressure sensor based on carbon nanotubes/epoxy composite under cyclic loading," *I2MTC 2018 - 2018 IEEE International Instrumentation and Measurement Technology Conference: Discovering New Horizons in Instrumentation and Measurement, Proceedings*, pp. 1–5, 2018.

[26] S. Takenawa, "A magnetic type tactile sensor using a two-dimensional array of inductors," *Proceedings - IEEE International Conference on Robotics and Automation*, no. 1, pp. 3295–3300, 2009.

[27] A. Kimoto and Y. Matsue, "A new multifunctional tactile sensor for detection of material hardness," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 4, pp. 1334–1339, 2011.

[28] S. Baglio, L. Cantelli, F. Giusa, and G. Muscato, "Intelligent prodder: Implementation of measurement methodologies for material recognition and classification with humanitarian demining applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 64, no. 8, pp. 2217–2226, 2015.

[29] N. Jamali and C. Sammut, "Majority voting: Material classification by tactile sensing using surface texture," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 508–521, 2011.

[30] I. Bandyopadhyaya, D. Babu, A. Kumar, and J. Roychowdhury, "Tactile sensing based softness classification using machine learning," *Souvenir of the 2014 IEEE International Advance Computing Conference, IACC 2014*, pp. 1231–1236, 2014.

[31] M. V. Liarokapis, B. Calli, A. J. Spiers, and A. M. Dollar, "Unplanned, model-free, single grasp object classification with underactuated hands and force sensors," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-December, no. August, pp. 5073–5080, 2015.

[32] J. Konstantinova, G. Cotugno, A. Stilli, Y. Noh, and K. Althoefer, "Object classification using hybrid fiber optical force/proximity sensor," *Proceedings of IEEE Sensors*, vol. 2017-December, pp. 1–3, 2017.

[33] R. Ahmadi, M. Packirisamy, J. Dargahi, and R. Cecere, "Discretely loaded beam-type optical fiber tactile sensor for tissue manipulation and palpation in minimally invasive robotic surgery," *IEEE Sensors Journal*, vol. 12, no. 1, pp. 22–32, 2012.

[34] N. F. Lepora and B. Ward-Cherrier, "Superresolution with an optical tactile sensor," *IEEE International Conference on Intelligent Robots and Systems*, vol. 2015-December, pp. 2686–2691, 2015.

[35] K. Zhang, C. Butler, Q. Yang, and Y. Lu, "A fiber optic sensor for the measurement of surface roughness and displacement using artificial neural networks," *IEEE Transactions on Instrumentation and Measurement*, vol. 46, no. 4, pp. 899–902, 1997.

[36] J. S. Heo, J. H. Chung, and J. J. Lee, "Tactile sensor arrays using fiber Bragg grating sensors," *Sensors and Actuators, A: Physical*, vol. 126, no. 2, pp. 312–327, 2006.

[37] L. H. Negri, A. S. Paterno, M. Muller, and J. L. Fabris, "Sparse Force Mapping System Based on Compressive Sensing," *IEEE Transactions on Instrumentation and Measurement*, vol. 66, no. 4, pp. 830–836, 2017.

[38] A. Massaro, F. Spano, A. Lay-Ekuakille, P. Cazzato, R. Cingolani, and A. Athanassiou, "Design and characterization of a nanocomposite pressure sensor implemented in a tactile robotic system," *IEEE Transactions on Instrumentation and Measurement*, vol. 60, no. 8, pp. 2967–2975, 2011.

[39] S. Tsuji, "A tactile and proximity sensor by optical and electrical measurement," *Proceedings of IEEE Sensors*, vol. 61, no. 12, pp. 3312–3317, 2012.

[40] F. De Chiara, S. Wang, and H. Liu, "Creating a soft tactile skin employing fluorescence based optical sensing," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3375–3381, 2020.

[41] A. Yamaguchi and C. G. Atkeson, "Recent progress in tactile sensing and sensors for robotic manipulation: can we turn tactile sensing into vision?" *Advanced Robotics*, vol. 33, no. 14, pp. 661–673, 2019.

[42] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: the rising trend of vision based measurement," *IEEE Instrumentation and Measurement Magazine*, vol. 17, no. 3, pp. 41–47, 2014.

[43] D. Bouget, M. Allan, D. Stoyanov, and P. Jannin, "Vision-based and marker-less surgical tool detection and tracking: a review of the literature," *Medical Image Analysis*, vol. 35, pp. 633–654, 2017.

[44] T. O. S. U. R. F. Honda Motor Co., Ltd., "Sign based human-machine interaction," 2007.

[45] R. Sanjeevi, R. Nagaraja, and B. Radha Krishnan, "Vision-based surface roughness accuracy prediction in the CNC milling process (Al6061) using ANN," *Materials Today: Proceedings*, vol. 2214, p. 7853, 2020.

[46] P.-J. Duh, Y.-C. Sung, L.-Y. F. Chiang, Y.-J. Chang, and K.-W. Chen, "V-Eye: A Vision-based Navigation System for the Visually Impaired," *IEEE Transactions on Multimedia*, 2020.

[47] K. Shimonomura, "Tactile image sensors employing camera: A review," *Sensors (Switzerland)*, vol. 19, no. 18, p. 3933, 2019.

[48] M. K. Johnson and E. H. Adelson, "Retrographic sensing for the measurement of surface texture and shape," in *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, CVPR Workshops 2009*, vol. 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE, 2009, pp. 1070–1077.

[49] M. Ohka, H. Kobayash, J. Takata, and Y. Mitsuya, "An experimental optical three-axis tactile sensor featured with hemispherical surface," *Nihon Kikai Gakkai Ronbunshu, C Hen/Transactions of the Japan Society of Mechanical Engineers, Part C*, vol. 74, no. 6, pp. 1477–1484, 2008.

[50] K. Tanie, K. Komoriya, M. Kaneko, S. Tachi, and A. Fujikawa, "High Resolution Tactile Sensor." in *Proc. of 4th Int. Conf. on Robot Vision and Sensory Controls*, vol. 251, 1984, pp. 251–260.

[51] H. Maekawa, K. Tanie, K. Komoriya, M. Kaneko, C. Horiguchi, and T. Sugawara, "Development of a finger-shaped tactile sensor and its evaluation by active touch," in *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2, 1992, pp. 1327–1334.

[52] K. Kamiyama, H. Kajimoto, N. Kawakami, and S. Tachi, "Evaluation of a vision-based tactile sensor," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2004, no. 2, pp. 1542–1547, 2004.

[53] B. Ward-Cherrier, N. Pestell, L. Cramphorn, B. Winstone, M. E. Giannaccini, J. Rossiter, and N. F. Lepora, "The TacTip Family: Soft Optical Tactile Sensors with 3D-Printed Biomimetic Morphologies," *Soft Robotics*, vol. 5, no. 2, pp. 216–227, 2018.

[54] K. Shimonomura, H. Nakashima, and K. Nozu, "Robotic grasp control with high-resolution combined tactile and proximity sensing," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2016-June, pp. 138–143, 2016.

[55] H. Nakashima, K. Kagawa, and K. Shimonomura, "Combined tactile and proximity sensor employing compound-eye camera," *ITE Transactions on Media Technology and Applications*, vol. 3, no. 4, pp. 227–233, 2015.

[56] K. Nozu and K. Shimonomura, "Robotic bolt insertion and tightening based on in-hand object localization and force sensing," in *IEEE/ASME International Conference on Advanced Intelligent Mechatronics, AIM*, vol. 2018-July, 2018, pp. 310–315.

[57] W. Yuan, R. Li, M. A. Srinivasan, and E. H. Adelson, "Measurement of shear and slip with a GelSight tactile sensor," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2015-June, no. June, pp. 304–311, 2015.

[58] C. Yu, L. Lindenroth, J. Hu, J. Back, G. Abrahams, and H. Liu, "A vision-based soft somatosensory system for distributed pressure and temperature sensing," *IEEE Robotics and Automation Letters*, vol. 5, no. 2, pp. 3323–3329, 2020.

[59] A. Asadian, M. R. Kermani, and R. V. Patel, "A novel force modeling scheme for needle insertion using multiple kalman filters," *IEEE Transactions on Instrumentation and Measurement*, vol. 61, no. 2, pp. 429–438, 2012.

[60] 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.

[61] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," pp. 1–25, 2019.

[62] T. Delbruck and P. Lichtsteiner, "Fast sensory motor control based on event-based hybrid neuromorphic- procedural system," *Proceedings - IEEE International Symposium on Circuits and Systems*, no. 80 cm, pp. 845–848, 2007.

[63] T. Delbruck and M. Lang, "Robotic goalie with 3 ms reaction time at 4% CPU load using event-based dynamic vision sensor," *Frontiers in Neuroscience*, vol. 7, no. 7 NOV, pp. 1–7, 2013.

[64] J. Conradt, M. Cook, R. Berner, P. Lichtsteiner, R. J. Douglas, and T. Delbruck, "A pencil balancing robot using a pair of AER dynamic vision sensors," *Proceedings - IEEE International Symposium on Circuits and Systems*, pp. 781–784, 2009.

[65] J. Conradt, R. Berner, M. Cook, and T. Delbruck, "An embedded AER dynamic vision sensor for low-latency pole balancing," *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, pp. 780–785, 2009.

[66] Y. Lu, Z. Xue, G. S. Xia, and L. Zhang, "A survey on vision-based UAV navigation," *Geo-Spatial Information Science*, vol. 21, no. 1, pp. 21–32, 2018.

[67] J. Janai, F. Guney, A. Behl, and A. Geiger, "Computer vision for autonomous vehicles: Problems, datasets and state-of-the-art," *arXiv*, pp. arXiv—-1704, 2017.

[68] D. Falanga, S. Kim, and D. Scaramuzza, "How Fast Is Too Fast? the Role of Perception Latency in High-Speed Sense and Avoid," *IEEE Robotics and Automation Letters*, vol. 4, no. 2, pp. 1884–1891, 2019.

[69] J. J. Hagenaars, F. Paredes-Vallés, S. M. Bohté, and G. C. De Croon, "Evolved Neuromorphic Control for High Speed Divergence-Based Landings of MAVs," *IEEE Robotics and Automation Letters*, vol. 5, no. 4, pp. 6239–6246, 2020.

[70] R. S. Dimitrova, M. Gehrig, D. Brescianini, and D. Scaramuzza, "Towards Low-Latency High-Bandwidth Control of Quadrotors using Event Cameras," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 4294–4300, 2020.

[71] A. I. Maqueda, A. Loquercio, G. Gallego, N. Garcia, and D. Scaramuzza, "Event-Based Vision Meets Deep Learning on Steering Prediction for Self-Driving Cars," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, no. Dl, pp. 5419–5427, 2018.

[72] G. Cohen, S. Afshar, G. Orchard, J. Tapson, R. Benosman, and A. Van Schaik, "Spatial and Temporal Downsampling in Event-Based Visual Classification," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 10, pp. 5030–5044, 2018.

[73] N. Khan, K. Iqbal, and M. G. Martini, "Lossless Compression of Data from Static and Mobile Dynamic Vision Sensors-Performance and Trade-Offs," *IEEE Access*, vol. 8, pp. 103 149–103 163, 2020.

[74] ——, "Time Aggregation based Lossless Video Encoding for Neuromorphic Vision Sensor Data," *IEEE Internet of Things Journal*, vol. 4662, no. c, 2020.

[75] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep learning*. MIT press Cambridge, 2016, vol. 1, no. 2.

[76] Z. Gong, P. Zhong, and W. Hu, "Diversity in Machine Learning," *IEEE Access*, vol. 7, pp. 64 323–64 350, 2019.

[77] P. Ciancarini, F. Poggi, and D. Russo, "Big Data Quality: A Roadmap for Open Data," in *Proceedings - 2016 IEEE 2nd International Conference on Big Data Computing Service and Applications, BigDataService 2016*, 2016, pp. 210–215.

[78] M. Fernández-Delgado, M. S. Sirsat, E. Cernadas, S. Alawadi, S. Barro, and M. Febrero-Bande, "An extensive experimental survey of regression methods," *Neural Networks*, vol. 111, pp. 11–34, 2019.

[79] R. Nason, P. Lloyd, and I. Ginns, "Format-free databases and the construction of knowledge in primary school science projects," *Research in Science Education*, vol. 26, no. 3, pp. 353–373, 1996.

[80] C. Lucas, O. Maier, and M. P. Heinrich, "Shallow fully-connected neural networks for ischemic stroke-lesion segmentation in MRI," in *Informatik aktuell*. Springer, 2017, pp. 261–266.

[81] E. Mingolla and D. Bullock, "Neurocomputing: Foundations of Research," in *Neural Networks*, J. A. Anderson and E. Rosenfeld, Eds. Cambridge, MA, USA: MIT Press, 1989, vol. 2, no. 5, ch. Learning R, pp. 405–409.

[82] J. Masci, U. Meier, D. Cireşan, and J. Schmidhuber, "Stacked convolutional auto-encoders for hierarchical feature extraction," in *International conference on artificial neural networks*. Springer, 2011, pp. 52–59.

[83] G. E. Hinton, S. Osindero, and Y.-W. Teh, "A fast learning algorithm for deep belief nets," *Neural computation*, vol. 18, no. 7, pp. 1527–1554, 2006.

[84] R. Salakhutdinov and G. Hinton, "Deep boltzmann machines," in *Artificial intelligence and statistics*, 2009, pp. 448–455.

[85] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in Neural Information Processing Systems*, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, vol. 3, no. January, pp. 2672–2680.

[86] M. A. Ponti, L. S. F. Ribeiro, T. S. Nazare, T. Bui, and J. Collomosse, "Everything you wanted to know about deep learning for computer vision but were afraid to ask," in *2017 30th SIBGRAPI Conference on Graphics, Patterns and Images Tutorials (SIBGRAPI-T)*, 2017, pp. 17–41.

[87] 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. 2278–2323, 1998.

[88] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Communications of the ACM*. Curran Associates, Inc., 2017, vol. 60, no. 6, pp. 84–90.

[89] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," apr 2015.

[90] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1–9, 2015.

[91] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2016-December, pp. 770–778, 2016.

[92] W. Liu, Z. Wang, X. Liu, N. Zeng, Y. Liu, and F. E. Alsaadi, "A survey of deep neural network architectures and their applications," *Neurocomputing*, vol. 234, pp. 11–26, 2017.

[93] 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, 2013.

[94] A. Waibel, T. Hanazawa, G. Hinton, K. Shikano, and K. J. Lang, "Phoneme Recognition Using Time-Delay Neural Networks," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 37, no. 3, pp. 328–339, 1989.

[95] A. Graves, A. R. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, no. 6, pp. 6645–6649, 2013.

[96] Y. Du, W. Wang, and L. Wang, "Hierarchical recurrent neural network for skeleton based action recognition," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 1110–1118, 2015.

[97] R. J. Frank, N. Davey, and S. P. Hunt, "Time series prediction and neural networks," *Journal of Intelligent and Robotic Systems: Theory and Applications*, vol. 31, no. 1-3, pp. 91–103, 2001.

[98] Y. Bengio, P. Simard, and P. Frasconi, "Learning Long-Term Dependencies with Gradient Descent is Difficult," *IEEE Transactions on Neural Networks*, vol. 5, no. 2, pp. 157–166, 1994.

[99] S. Hochreiter and J. Schmidhuber, "Long Short-Term Memory," *Neural Computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[100] E. Stromatias, M. Soto, T. Serrano-Gotarredona, and B. Linares-Barranco, "An event-driven classifier for spiking neural networks fed with synthetic or dynamic vision sensor data," *Frontiers in Neuroscience*, vol. 11, no. JUN, pp. 1–17, 2017.

[101] J. Schmidhuber, "Deep Learning in neural networks: An overview," *Neural Networks*, vol. 61, pp. 85–117, 2015.

[102] F. Perez-Peña, A. Morgado-Estevez, A. Linares-Barranco, A. Jimenez-Fernandez, F. Gomez-Rodriguez, G. Jimenez-Moreno, and J. Lopez-Coronado, "Neuro-inspired spike-based motion: From dynamic vision sensor to robot motor open-loop control through spike-VITE," *Sensors (Switzerland)*, vol. 13, no. 11, pp. 15 805–15 832, 2013.

[103] J. L. Lobo, J. Del Ser, A. Bifet, and N. Kasabov, "Spiking neural networks and online learning: An overview and perspectives," *arXiv*, vol. 121, pp. 88–100, 2019.

[104] M. Seeger, "Gaussian processes for machine learning." *International journal of neural systems*, vol. 14, no. 2, pp. 69–106, 2004.

[105] Y. Sui and L. Zhang, "Visual tracking via locally structured Gaussian process regression," *IEEE Signal Processing Letters*, vol. 22, no. 9, pp. 1331–1335, 2015.

[106] S. C.-X. Li and B. M. Marlin, "A scalable end-to-end gaussian process adapter for irregularly sampled time series classification," in *Advances in neural information processing systems*, 2016, pp. 1804–1812.

[107] A. C. Damianou and N. D. Lawrence, "Deep Gaussian processes," in *Journal of Machine Learning Research*, vol. 31, no. 3, 2013, pp. 207–215.

[108] R. B. Gramacy and H. K. Lee, "Bayesian treed Gaussian process models with an application to computer modeling," *Journal of the American Statistical Association*, vol. 103, no. 483, pp. 1119–1130, 2008.

[109] S. Shirmohammadi and A. Ferrero, "Camera as the instrument: the rising trend of vision based measurement," *IEEE Instrumentation and Measurement Magazine*, vol. 17, no. 3, pp. 41–47, 2014.

[110] A. Ikeda, Y. Kurita, J. Ueda, Y. Matsumoto, and T. Ogasawara, "Grip force control for an elastic finger using vision-based incipient slip feedback," *2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, vol. 1, pp. 810–815, 2004.

[111] K. Vlack, T. Mizota, N. Kawakami, K. Kamiyama, H. Kajimoto, and S. Tachi, "GelForce: A vision-based traction field computer interface," *Conference on Human Factors in Computing Systems - Proceedings*, vol. 2, pp. 1154–1155, 2005.

[112] D. L. James and D. K. Pai, "Multiresolution green's function methods for interactive simulation of large-scale elastostatic objects," *ACM Transactions on Graphics*, vol. 22, no. 1, pp. 47–82, 2003.

[113] G. Obinata, A. Dutta, N. Watanabe, and N. Moriyam, "Vision Based Tactile Sensor Using Transparent Elastic Fingertip for Dexterous Handling," *Mobile Robots: Perception & Navigation*, no. February, pp. 137–148, 2007.

[114] Y. Ito, Y. W. Kim, and G. Obinata, "Slippage degree estimation by using vision-based tactile sensor for dexterous handling," *IFAC Proceedings Volumes (IFAC-PapersOnline)*, vol. 42, no. 16, pp. 281–286, 2009.

[115] Y. Ito, Y. Kim, and G. Obinata, "Contact region estimation based on a vision-based tactile sensor using a deformable touchpad," *Sensors (Switzerland)*, vol. 14, no. 4, pp. 5805–5822, 2014.

[116] K. Zhao, X. Li, C. Lu, G. Lu, and Y. Wang, "Video-based slip sensor for multidimensional information detecting in deformable object grasp," *Robotics and Autonomous Systems*, vol. 91, pp. 71–82, 2017.

[117] A. Kolker, M. Jokesch, and U. Thomas, "An optical tactile sensor for measuring force values and directions for several soft and rigid contacts," *47th International Symposium on Robotics, ISR 2016*, vol. 2016, pp. 63–68, 2016.

[118] K. Kamiyama, K. Vlack, T. Mizota, H. Kajimoto, N. Kawakami, and S. Tachi, "Vision-based sensor for real-time measuring of surface traction fields," *IEEE Computer Graphics and Applications*, vol. 25, no. 1, pp. 68–75, 2005.

[119] Y. Ito, Y. Kim, and G. Obinata, "Robust slippage degree estimation based on reference update of vision-based tactile sensor," *IEEE Sensors Journal*, vol. 11, no. 9, pp. 2037–2047, 2011.

[120] B. Winstone, G. Griffiths, C. Melhuish, T. Pipe, and J. Rossiter, "TACTIP - Tactile fingertip device, challenges in reduction of size to ready for robot hand integration," in *2012 IEEE International Conference on Robotics and Biomimetics, ROBIO 2012 - Conference Digest*. IEEE, 2012, pp. 160–166.

[121] D. Ma, E. Donlon, S. Dong, and A. Rodriguez, "Dense tactile force estimation using gelslim and inverse FEM," *Proceedings - IEEE International Conference on Robotics and Automation*, vol. 2019-May, pp. 5418–5424, 2019.

[122] A. I. Aviles, A. Marban, P. Sobrevilla, J. Fernandez, and A. Casals, "A recurrent neural network approach for 3D vision-based force estimation," *2014 4th International Conference on Image Processing Theory, Tools and Applications, IPTA 2014*, pp. 1–6, 2015.

[123] A. I. Aviles, S. Alsaleh, P. Sobrevilla, and A. Casals, "Sensorless force estimation using a neuro-vision-based approach for robotic-assisted surgery," *International IEEE/EMBS Conference on Neural Engineering, NER*, vol. 2015-July, pp. 86–89, 2015.

[124] M. Ikeda, J. Moriguchi, S. Sakuragi, and F. Ohashi, "Association of past diseases with levels of cadmium and tubular dysfunction markers in urine of adult women in non-polluted areas in Japan," *International Archives of Occupational and Environmental Health*, vol. 86, no. 3, pp. 343–355, 2013.

[125] N. Alt and E. Steinbach, "Navigation and manipulation planning using a visuo-haptic sensor on a mobile platform," *IEEE Transactions on Instrumentation and Measurement*, vol. 63, no. 11, pp. 2570–2582, 2014.

[126] N. Haouchine, W. Kuang, S. Cotin, and M. Yip, "Vision-Based Force Feedback Estimation for Robot-Assisted Surgery Using Instrument-Constrained Biomechanical Three-Dimensional Maps," *IEEE Robotics and Automation Letters*, vol. 3, no. 3, pp. 2160–2165, 2018.

[127] S. Luo, W. Yuan, E. Adelson, A. G. Cohn, and R. Fuentes, "ViTac: Feature Sharing between Vision and Tactile Sensing for Cloth Texture Recognition," *Proceedings - IEEE International Conference on Robotics and Automation*, pp. 2722–2727, 2018.

[128] A. J. Spiers, M. V. Liarokapis, B. Calli, and A. M. Dollar, "Single-Grasp Object Clas-

sification and Feature Extraction with Simple Robot Hands and Tactile Sensors," *IEEE Transactions on Haptics*, vol. 9, no. 2, pp. 207–220, 2016.

[129] O. Kroemer, C. H. Lampert, and J. Peters, "Learning dynamic tactile sensing with robust vision-based training," *IEEE Transactions on Robotics*, vol. 27, no. 3, pp. 545–557, 2011.

[130] A. Rigi, F. B. Naeini, D. Makris, and Y. Zweiri, "A novel event-based incipient slip detection using dynamic active-pixel vision sensor (DAVIS)," *Sensors (Switzerland)*, vol. 18, no. 2, p. 333, jan 2018.

[131] M. Yang, S. C. Liu, and T. Delbruck, "As dynamic vision sensor with 1% temporal contrast sensitivity and in-pixel asynchronous delta modulator for event encoding," *IEEE Journal of Solid-State Circuits*, vol. 50, no. 9, pp. 2149–2160, 2015.

[132] M. A. Hopcroft, W. D. Nix, and T. W. Kenny, "What is the Young's modulus of silicon?" *Journal of Microelectromechanical Systems*, vol. 19, no. 2, pp. 229–238, 2010.

[133] Y. E. Shao and S. C. Lin, "Using a time delay neural network approach to diagnose the out-of-control signals for a multivariate normal process with variance shifts," *Mathematics*, vol. 7, no. 10, p. 959, 2019.

[134] I. Mukherjee and S. Routroy, "Comparing the performance of neural networks developed by using Levenberg-Marquardt and Quasi-Newton with the gradient descent algorithm for modelling a multiple response grinding process," *Expert Systems with Applications*, vol. 39, no. 3, pp. 2397–2407, 2012.

[135] K. Levenberg, "A method for the solution of certain non-linear problems in least squares," *Quarterly of Applied Mathematics*, vol. 2, no. 2, pp. 164–168, 1944.

[136] R. M. Neal, *Bayesian Learning for Neural Networks*. Springer Science & Business Media, 1996, vol. 1, no. 118.

[137] W. Yuan, S. Dong, and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force," *Sensors (Switzerland)*, vol. 17, no. 12, 2017.

[138] C. Sferrazza, A. Wahlsten, C. Trueeb, and R. D'Andrea, "Ground truth force distribution for learning-based tactile sensing: a finite element approach," *arXiv*, vol. 7, pp. 173 438–173 449, 2019.

[139] W. Yuan, C. Zhu, A. Owens, M. A. Srinivasan, and E. H. Adelson, "Shape-independent hardness estimation using deep learning and a GelSight tactile sensor," *arXiv*, pp. 951–958, 2017.

[140] K. Kumagai and K. Shimonomura, "Event-based Tactile Image Sensor for Detecting Spatio-Temporal Fast Phenomena in Contacts," *2019 IEEE World Haptics Conference, WHC 2019*, no. Fa Ii, pp. 343–348, 2019.

[141] X. Huang, R. Muthusamy, E. Hassan, Z. Niu, L. Seneviratne, D. Gan, and Y. Zweiri, "Neuromorphic vision based contact-level classification in robotic grasping applications," *Sensors (Switzerland)*, vol. 20, no. 17, pp. 1–15, 2020.

[142] R. Muthusamy, X. Huang, Y. Zweiri, L. Seneviratne, D. Gan, and R. Muthusamy, "Neuromorphic Event-Based Slip Detection and Suppression in Robotic Grasping and Manipulation," pp. 153 364–153 384, 2020.

[143] B. Ward-Cherrier, N. Pestell, and N. F. Lepora, "NeuroTac: A Neuromorphic Optical Tactile Sensor applied to Texture Recognition," *arXiv*, 2020.

[144] F. Baghaei Naeini, A. M. Alali, R. Al-Husari, A. Rigi, M. K. Al-Sharman, D. Makris, and Y. Zweiri, "A Novel Dynamic-Vision-Based Approach for Tactile Sensing Applications," *IEEE Transactions on Instrumentation and Measurement*, vol. 69, no. 5, pp. 1881–1893, 2020.

[145] F. B. Naeini, D. Makris, D. Gan, and Y. Zweiri, "Dynamic-vision-based force measurements using convolutional recurrent neural networks," *Sensors (Switzerland)*, vol. 20, no. 16, pp. 1–15, 2020.

[146] X. Shi, Z. Chen, H. Wang, D. Y. Yeung, W. K. Wong, and W. C. Woo, "Convolutional LSTM network: A machine learning approach for precipitation nowcasting," *Advances in Neural Information Processing Systems*, vol. 2015-January, pp. 802–810, 2015.

[147] J. Y. H. Ng, M. Hausknecht, S. Vijayanarasimhan, O. Vinyals, R. Monga, and G. Toderici, "Beyond short snippets: Deep networks for video classification," *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 07-12-June-2015, pp. 4694–4702, 2015.

[148] F. Chollet, "Keras," 2016. [Online]. Available: http://keras.io

[149] M. Henaff, A. Szlam, and Y. Lecun, "Recurrent orthogonal networks and long-memory tasks," *33rd International Conference on Machine Learning, ICML 2016*, vol. 5, pp. 2978–2986, 2016.

[150] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[151] C. Shorten and T. M. Khoshgoftaar, "A survey on Image Data Augmentation for Deep Learning," *Journal of Big Data*, vol. 6, no. 1, 2019.

[152] A. Mikołajczyk and M. Grochowski, "Data augmentation for improving deep learning in image classification problem," in *2018 International Interdisciplinary PhD Workshop, IIPhDW 2018*. IEEE, 2018, pp. 117–122.

[153] D. Yasmina, R. Karima, and A. Ouahiba, "Traffic signs recognition with deep learning," in *Proceedings of the 2018 International Conference on Applied Smart Systems, ICASS 2018*, 2019, pp. 1–5.

[154] I. Sato, H. Nishimura, and K. Yokoi, "APAC: Augmented PAttern Classification with Neural Networks," *arXiv preprint arXiv:1505.03229*, 2015.

[155] E. D. Cubuk, B. Zoph, D. Mane, V. Vasudevan, and Q. V. Le, "Autoaugment: Learning augmentation strategies from data," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2019-June, 2019, pp. 113–123.

[156] H. Bagherinezhad, M. Horton, M. Rastegari, and A. Farhadi, "Label Refinery: Improving ImageNet Classification through Label Progression," *arXiv preprint arXiv:1805.02641*, 2018.

[157] K. J. Piczak, "Environmental sound classification with convolutional neural networks," in *IEEE International Workshop on Machine Learning for Signal Processing, MLSP*, vol. 2015-November. IEEE, 2015, pp. 1–6.

[158] A. L. Guennec, S. Malinowski, and R. Tavenard, "Data Augmentation for Time Series Classification using Convolutional Neural Networks," in *ECML/PKDD Workshop on Advanced Analytics and Learning on Temporal Data*, 2016.

[159] X. Yi, E. Walia, and P. Babyn, "Generative adversarial network in medical imaging: A review," *Medical Image Analysis*, vol. 58, p. 101552, 2019.

[160] J. Y. Zhu, P. Krähenbühl, E. Shechtman, and A. A. Efros, "Generative visual manipulation on the natural image manifold," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 9909 LNCS. Springer, 2016, pp. 597–613.

[161] E. Denton, S. Chintala, A. Szlam, and R. Fergus, "Deep generative image models using a laplacian pyramid of adversarial networks," in *Advances in Neural Information Processing Systems*, C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, Eds. Curran Associates, Inc., 2015, vol. 2015-January, pp. 1486–1494.

[162] A. Radford, L. Metz, and S. Chintala, "Unsupervised representation learning with deep convolutional generative adversarial networks," *4th International Conference on Learning Representations, ICLR 2016 - Conference Track Proceedings*, 2016.

[163] O. Mogren, "C-RNN-GAN: Continuous recurrent neural networks with adversarial training," *arXiv preprint arXiv:1611.09904*, 2016.

[164] S. L. Hyland, C. Esteban, and G. Rätsch, "Real-valued (medical) time series generation with recurrent conditional GANs," *arXiv*, 2017.

[165] G. Ramponi, P. Protopapas, M. Brambilla, and R. Janssen, "T-CGAN: Conditional generative adversarial network for data augmentation in noisy time series with irregular sampling," *arXiv*, 2018.

[166] J. Yoon, D. Jarrett, and M. van der Schaar, "Time-series Generative Adversarial Networks," in *Advances in Neural Information Processing Systems 32*, H. Wallach, H. Larochelle, A. Beygelzimer, F. d\textquotesingle Alché-Buc, E. Fox, and R. Garnett, Eds. Curran Associates, Inc., 2019, pp. 5508–5518.

[167] J. Wang and L. Perez, "The effectiveness of data augmentation in image classification using deep learning," *arXiv*, 2017.

[168] Z. Hussain, F. Gimenez, D. Yi, and D. Rubin, "Differential Data Augmentation Techniques for Medical Imaging Classification Tasks," in *AMIA ... Annual Symposium proceedings. AMIA Symposium*, vol. 2017. American Medical Informatics Association, 2017, pp. 979–984.

[169] E. Mueggler, H. Rebecq, G. Gallego, T. Delbruck, and D. Scaramuzza, "The event-camera dataset and simulator: Event-based data for pose estimation, visual odometry, and slam," *The International Journal of Robotics Research*, vol. 36, no. 2, pp. 142–149, 2017.

[170] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[171] 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, 2017.

[172] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *arXiv*, 2017, pp. 5998–6008.

[173] C. Sferrazza, T. Bi, and R. D'Andrea, "Learning the sense of touch in simulation: a sim-to-real strategy for vision-based tactile sensing," *arXiv*, 2020.