VISUAL SONAR: ESTIMATING DEPTH USING SOUND WAVES AND NEUROMORPHIC CAMERAS

by

Abhijith R Bagepalli

A thesis submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Master of Science in
Electrical Engineering

Charlotte

2020

Approved by:

_____
Dr. Andrew Willis

_____
Dr. Chen Chen

_____
Dr. Thomas Weldon

ABSTRACT

ABHIJITH R BAGEPALLI. Visual sonar: Estimating depth using sound waves and neuromorphic cameras. (Under the direction of DR. ANDREW WILLIS)

This thesis proposes a novel method to recover depth of scene objects using an acoustic source and a calibrated neuromorphic camera or event camera sensor. The proposed system is a non-contact, monocular depth estimation method that observes subtle mechanical vibrations in scene objects induced by sound waves and uses geometric image formation models to recover depth.

Neuromorphic cameras are high speed cameras that are capable of capturing subtle motion and vibrations on the surface of scene objects caused by sound waves. The neuromorphic camera observes a change in intensity at every pixel which triggers an asynchronous output of an event characterized by its pixel coordinates, (x, y), polarity (p) (i.e positive or negative change in intensity) and the time stamp (t_s). Using this event data in conjunction with the geometric setup of the optical system, we recover depth by estimating the time of flight for an emitted sound wave to strike an object's surface. The method proposed in this thesis to estimate depth using an acoustic excitation signal and a neuromorphic camera is the first of its kind.

Experiments were conducted by subjecting a sheet of paper to an impulse-like sound wave and a sinusoidal sound wave. The results show that the proposed method is able to estimate depth with an error of $\pm 1cm$. Further, we demonstrate how we can reconstruct a sinusoidal excitation signal that was emitted by analyzing the vibrations of the scene object and estimating the signal's frequency. Results indicate that our proposed method is able to estimate the signal's frequency with an error of 2.2 Hz. The proposed acoustic-optical sensing mechanism shows potential uses cases in estimating the structural properties of the object, vibration analysis, robotics, etc.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor Dr. Andrew Willis, without whom this thesis would not be possible. I would like to thank Dr. Willis for his support, guidance, and for giving me the opportunity to work under him. He has my eternal gratitude for his time and immense patience with me. He has helped in my academic and professional growth and I have been extremely fortunate to be his student. I am also grateful to Dr. Chen Chen and Dr. Thomas Weldon for serving on my thesis committee and for their valuable suggestions and feedback.

I would like to express my appreciation to my colleagues in the Machine Vision Laboratory, Jincheng Zhang and Sajjad Hossain for their support and exchange of ideas which have aided in a deeper understanding of computer vision.

Lastly, I would like to thank my family for their unwavering support through out my journey as a graduate student.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Estimating depth from 2D images is a crucial step in tasks such as scene reconstruction, 3D object recognition, segmentation, and detection. The problem of depth estimation can be defined as the derivation of the distance from the camera to each point of the scene in a 2D image. The human visual system perceives the depth of objects in front of us and makes sense of our 3D world with ease. However, depth perception and the ability to understand the 3D structure of scene objects are still complex tasks for computers to perform accurately and quickly. These tasks play an important role in computer vision and computer graphics and have numerous applications such as robot navigation, 3D modeling, autonomous driving, medical imaging and structural engineering [1, 2].

Prior work on 3D reconstruction have focused on methods that use multiple images such as stereo vision [1] and rely on triangulation to estimate depth. These algorithms tend to be inaccurate when the baseline distance between the two camera positions is large. They also tend to fail for texture-less regions where correspondences cannot be reliably found [3]. Furthermore, these methods cannot be used when only a single image is available.

In more recent years, monocular depth estimation algorithms have gained traction as stereo reconstruction methods require more resources, complex physical setups and large amounts of data when compared to monocular depth estimation. Monocular cues such as texture variations, texture gradients, light, and shading can provide useful depth and 3D information and hence several authors [3, 4, 5, 6] have developed methods for depth estimation from a single image.

Data acquisition for depth estimation and 3D reconstruction can occur from a

multitude of methods including 2D images using visible light, infrared light used by RGB-D cameras such as the Microsoft Kinect [7] or radio waves used in Synthetic Aperture Radar(SAR) [8]. Sound waves is another type of excitation signal that is used to estimate depth in methods such as SONAR [9], ultrasonic techniques for 3D reconstruction and 3D room reconstruction from sound [10, 11, 12].

This thesis proposes the use of an acoustic-optical method for depth estimation. The proposed method uses an acoustic source and a calibrated neuromorphic camera to estimate the depth of scene objects. When an object is subjected to sound waves, the vibrations on the surface are typically hard to see with the naked eye and traditional CMOS based cameras. However, recently introduced event cameras or Dynamic Vision Sensor (DVS) cameras [13] are capable of sensing changes in the observed intensities at a pixel level with each event sampled in the order of microseconds per pixel. This allows us to sense and capture subtle motion in scenery such as vibrations on an object's surface caused by sound waves.

1.1 Neuromorphic cameras

Neuromorphic cameras or event cameras [13] are bio-inspired imaging sensors which sense at the pixel-level and output only up/down changes in the observed intensities as continuous stream of pixel events rather than image frames. This architecture has a number of benefits including:

- High dynamic range since exposure for each pixel is independently adjusted.
- Asynchronous pixel event data streams which reduces redundant information thus resulting in increased update rates by several orders of magnitude which in turn reduces CPU usage and, by extension, CPU power consumption [14, 15].
- The event data streams have latencies of $\sim$12-15 μsec allowing nearly continuous visual control loops [16, 17].

- The events generated are time stamped which is accurate within 1 μsec providing unprecedented per pixel temporal resolution.

1.2 Overview

The method proposed in this thesis observes subtle mechanical vibrations in scene objects induced by sound waves and estimates the Time of Flight (ToF) for an emitted acoustic wave to hit the surface of an object using the events observed by an event camera. The ToF is used in conjunction with the geometric setup of the optical system, which is used to derive the calculation for depth.

An overview of the steps involved in the methodology is shown below:

1. Generate acoustic excitation waves.
2. Calibrate the camera to obtain estimates of the intrinsic and extrinsic parameters for optical image formation.
3. Estimate time instance at which sound is emitted.
4. Estimate time instance at which sound waves strike the object.
5. Estimate time-of-flight for the acoustic wave.
6. Estimate the speaker-to-object distance using the time of flight.
7. Estimate the object depth using the geometric model for the system's optical image formation and the estimated speaker-to-object distance.

1.3 Contribution

This thesis proposes a novel non contact, monocular depth estimation method using an acoustic source and a calibrated neuromorphic sensor. This acoustic-optical setup using a neuromorphic sensor is a new area in the existing literature and thus a significant intellectual merit of this thesis is the development of mathematical models

and experimental methods to estimate depth using the event data generated by the neuromorphic sensor. We propose a geometrical setup for the optical system and derive an equation to estimate depth using this geometry and the output generated by the event camera.

Additionally, we demonstrate that it is possible to recover the frequency of a sinusoidal excitation signal by analyzing the vibrations in the scene by fitting a sinusoid to the magnitude of motion in the image. Recovering the sound signal emitted and modeling the vibrations on the object's surface shows potential to estimate the structural properties of an object akin to vibration analysis [18, 19, 20] and modal analysis [21, 22], employed in structural and civil engineering.

1.4 Outline

This thesis is outlined as follows: Chapter 2 presents background information and concepts that apply to subsequent chapters such as a taxonomy of the existing data acquisition methods, the pinhole camera model for image formation and the physics of acoustic wave propagation. Chapter 3 is dedicated to neuromorphic cameras. It introduces the novel sensor, compares them with traditional frame based cameras and, describes in detail the principle of operation and how data from these sensors are represented. In Chapter 4, the methodology used in this thesis is described in detail and lays down the theory of all the components involved in the methodology. We provide specifications of the experimental setup used and present the results of the experiments conducted in Chapter 5. Finally, we summarize the work done as a conclusion in Chapter 6.

CHAPTER 2: BACKGROUND

This chapter provides background information and describes concepts that are applied in the subsequent chapters. Topics include existing data acquisition methods used for depth estimation and 3D reconstruction, neuromorphic image sensors, acoustic wave propagation, camera model and camera parameters.

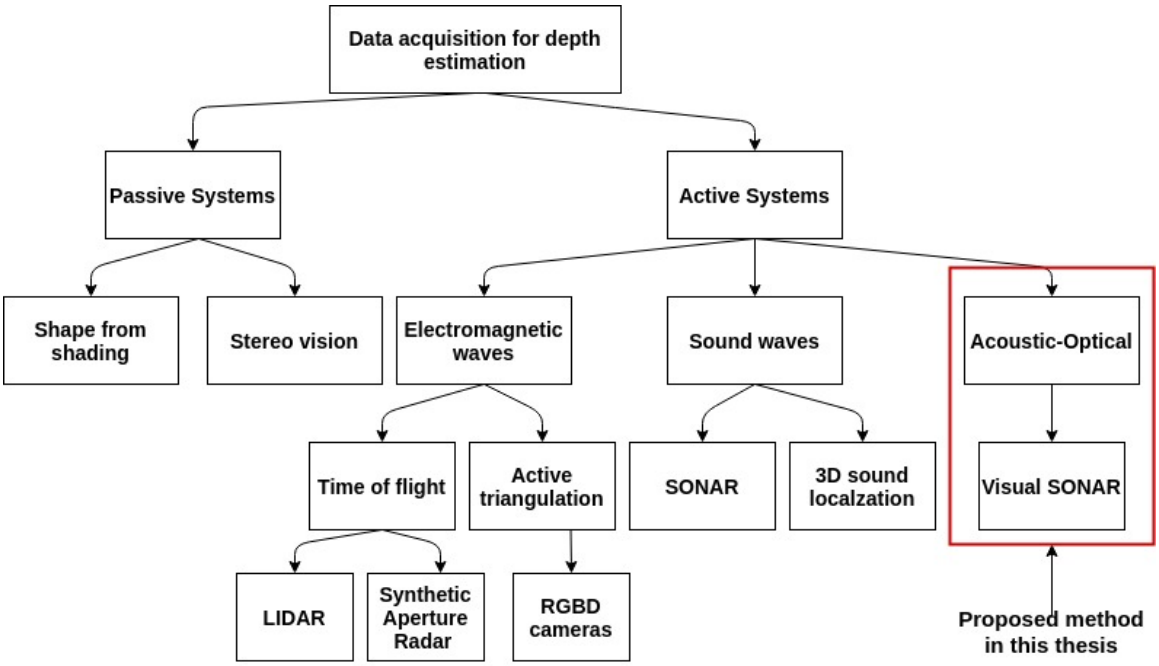


Figure 2.1: Taxonomy of data acquisition methods for depth estimation

Data acquisition for 3D reconstruction and depth estimation can be broadly classified into two categories namely, passive and active systems. In passive systems, the acquisition is done without any interaction with the object whereas active systems acquire data by interacting with the object using direct contact or a projection of some form of energy onto the object. The taxonomy of these systems are shown in Figure 2.1.

2.1 Passive systems

Passive systems do not use any external light and only use the ambient light to determine the 3D structure of an object. Examples of passive systems include shape from shading [23] and stereo vision.

- Shape from shading: This system infers the 3D shape of an object's surface from one image of the object using its shading information [23]. It uses variations in the brightness to recover the shape of the object.

- Stereo vision system:

  - In a stereo vision system, the 3D structure of an object is obtained using two or more images, each acquired from different viewpoints. The depth information in this case, is obtained in the form of a disparity map.

  - In a two camera system or binocular stereo system, given two images acquired from different viewpoints, stereo matching algorithms identify the corresponding points in both the images related to the same scene.

  - Knowing these correspondences and the camera geometry, the 3D world co-ordinates can be reconstructed through triangulation [24]. Triangulation is the process of determining the location of a 3D point given its projections onto two or more images.

2.2 Active systems

Active sensors operate by projecting energy wave or excitation signal on an object and then analyzing the transmitted or reflected wave. Active systems are subdivided based on the type of wave projected: electromagnetic waves and acoustic waves.

2.2.1 Electromagnetic waves

These systems use waves that lie in the electromagnetic spectrum such as visible light, infrared light or radio waves (Radar). The measurement principles used are

active triangulation or active stereo and time of flight.

- Active triangulation/stereo:

  - Active triangulation or active stereo vision systems are similar to the passive stereo vision systems but have an additional light source such as a laser or an infrared light pattern (structured light) which is projected onto the scene.

  - The cameras in the system detect this pattern and using the difference between the known pattern and the detected pattern, depth is calculated using triangulation.

  - Active stereo systems are useful in regions where there is a lack of light and/or texture on the object. The infrared projector or another light source will flood the scene with texture which reduces the dependency of an external light source.

  - A number of common RGB-D sensors utilise the structured light technology including the Microsoft Kinect [7], the Asus Xtion [25], and the Orbbec Astra [26].

- Time of Flight(ToF):

  - These systems measure the time that a wave emitted by a transmitter unit requires to travel to an object and back to a detector.

  - Examples of ToF systems are LIDAR where the wave emitted is a laser beam and Synthetic Aperture Radar(SAR) [8], a form of imaging radar where the wave emitted is a Radio wave.

2.2.2 Acoustic waves

This category of systems emit acoustic waves for 3D reconstruction and 3D sound localization. Examples include the imaging Sonar or synthetic aperture sonar(SAS) [10] which uses ultrasound signals whose frequencies are above 20kHz. SAS combines many acoustic pings to form an image with much higher resolution than conventional

sonars. SAS has useful applications in marine research, underwater construction work, offshore oil and gas, and in the military sector.

Another application of an acoustic wave system is 3D sound localization for surveillance applications where humans or cameras have no direct line of sight with the sound sources. In such cases, the ability to estimate the direction of the sources of danger relying on sound becomes extremely important.

The localization model in [11] is based on a neuromorphic microphone that takes advantage of the biologically-based monaural spectral cues to localize sound sources in a plane. The authors of [12] proposed a method to reconstruct the 3D structure (sensing the shape) of generic convex rooms from acoustic signals. They achieve microphones and sources localization and wall estimation by calculating the Time of arrival (TOA) of the direct path and echoes of the signal from the source to the microphone.

2.2.3 Acoustic-Optical

Based on the existing literature on the taxonomy of data acquisition methods for 3D reconstruction, the method proposed in this thesis work can be classified into a third category of active sensors which uses images and acoustic waves to estimate the depth of an object.

The proposed sensor setup directs acoustic signals at an object's surface. Next, the vibrations on the object's surface are captured using calibrated neuromorphic cameras. Using the time stamps generated by the camera, we estimate the Time of Flight (ToF) for an emitted acoustic wave to hit the surface of an object. The ToF is then used in conjunction with the geometric setup of the optical system, which is used to derive the calculation for depth.

2.3 Pinhole camera model

The pinhole camera model provides a mathematical relationship between the co-ordinates of a 3D point in the world coordinate space and its projection onto the 2D image plane. It describes a camera with a pinhole aperture and image plane as shown in Figure 2.2.

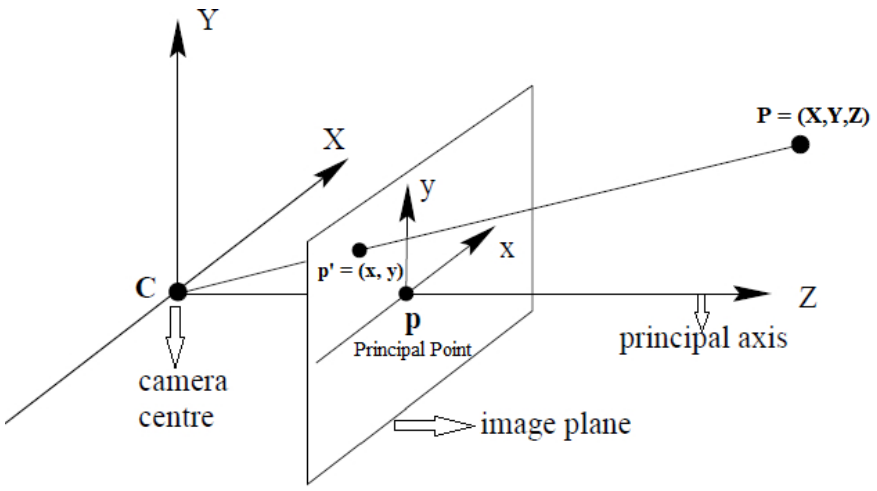


Figure 2.2: Pinhole camera model showing the relationship between 3D world points and their projection onto the image plane.

The centre of projection, C, is referred to as the camera centre or optical centre and is the origin of the Euclidean coordinate system. The plane $Z = f$ is called the image plane or focal plane. The line from the camera centre perpendicular to the image plane is called the principal axis. The point where the principal axis meets the image plane is called the principal point. The distance between the image plane and the camera centre is the focal length, f.

The pinhole model provides a projection mapping between 3D world points $P = [XYZ]^T$, and 2D points on the image plane $p' = [xy]^T$. It defines this perspective projection in terms of a set of parameters intrinsic to the camera like focal length, f in meters, *principal point* (c_x, c_y), in pixels and pixel size (s_x, s_y), in meters/pixel. The simplifications $f_x = \frac{f}{s_x}$ and $f_y = \frac{f}{s_y}$ are often made to express the focal length in units

of pixels. Using these equations, the resulting 2D pixel coordinates after projection can found using:

$$\begin{aligned} x &= f_x \frac{X}{Z} + c_x \\ y &= f_y \frac{Y}{Z} + c_y \end{aligned} \tag{2.1}$$

If the distance from the image plane (Z) is known for a given pixel value, the associated 3D point can be reconstructed using the inverse mapping:

$$\begin{aligned} X &= \frac{Z}{f_x}(x - c_x) \\ Y &= \frac{Z}{f_y}(y - c_y) \end{aligned} \tag{2.2}$$

2.3.1 Intrinsic Parameters

The pinhole camera model discussed above provides a relationship between points in 3D space to their corresponding 2D image pixel locations in terms of a set of parameters intrinsic to the camera. These intrinsic parameters can be expressed in a matrix form as:

$$K = \begin{bmatrix} \frac{f}{s_X} & 0 & c_x \\ 0 & \frac{f}{s_y} & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2.3}$$

The perspective projection of the 3D point (X, Y, Z) to the image point (x, y) can be expressed as:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_X \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{2.4}$$

2.3.2 Extrinsic Parameters

The extrinsic camera parameters are a set of geometric parameters used to determine accurately the fixed transformation between the camera frame and the world frame. Unlike with intrinsic parameters, the world origin is no longer located at the camera focal point as the camera is allowed to translate and rotate in space. The extrinsic parameters are now expressed in terms of a rotation matrix $R_{3\times3}$ and a translation vector $t_{3\times1}$. Thus the relationship between image coordinates and the world coordinates can be expressed as:

$$P_{image} = K \begin{bmatrix} R \mid t \end{bmatrix} P_{world} \tag{2.5}$$

or:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \tag{2.6}$$

2.4 Cameras

For the task of depth estimation, RGB-D cameras are typically used. They are popular low cost, high performance that combine a traditional color camera with an

infrared depth sensor to produce full HD color and range images at real-time frame rates. A common class of RGB-D sensors is those that leverage structured light approaches for depth estimation. Structured light sensors are active sensors that project a known infrared pattern out into the scene. These sensors are also equipped with infrared cameras located at a known baseline from the projector, which detect the pattern. Using the difference between the known pattern and the detected pattern, depth is calculated using triangulation.

2.4.1 Neuromorphic Cameras

In this thesis, we use a high frame rate, event camera or a neuromorphic camera [13]. Neuromorphic cameras are dynamic vision sensors that responds to changes in brightness. They do not capture images using shutters the way traditional cameras do which have a common exposure time. Instead, each pixel in the event camera independently measures changes in brightness/intensity as a continuous stream of pixel events.

Event cameras are used in applications such as object tracking [27, 28, 29, 30, 31], surveillance and monitoring [32, 33] where the high speed motion of scene objects causes changes in intensity which in turn leads to events generated by the event camera. Objects of interest are identified as a coherent event activity of neighboring pixels within a time window. Additionally, event cameras are used in applications such as object recognition when the objects are in motion [34, 35, 36, 37] and action recognition [16, 38]. These methods take a temporal approach to object recognition and utilize the precise timing information inherently present in the output of these biologically inspired sensors typically by using deep learning architectures.

Prior work related to the application of depth estimation - the topic covered in this thesis, use complex multi camera setup and visible light [39, 40, 41, 42, 43, 44] to capture depth and structural information on scene objects. The method proposed in this thesis differs from these existing depth estimation methods as it uses a single

camera setup with a novel acoustic-optical sensing mechanism to estimate depth.

The benefits of the neuromorphic or event cameras have led to an increase in their popularity in both academic and commercial worlds. Examples of commercially available event cameras include Samsung's DVS [45] and DAVIS (Dynamic and Active-Pixel Vision Sensor) by iniVation [46]. A more detailed comparison of the different types of sensors and their working is described in Chapter 3.

2.5 Acoustic wave propagation

An acoustic wave is a vibration that typically propagates as an audible wave of pressure, through a transmission medium such as a gas, liquid or solid. In terms of human physiology, sound is the reception of such waves and their perception by the brain. When the frequency of the wave lies between about 20 Hz and 20 kHz, it is in the human audible range. Sound waves above 20 kHz are known as ultrasound and are not perceptible by humans while sound waves below 20 Hz are known as infra-sound.

Figure 2.3: Frequency ranges of sound.

2.5.1 Physics of sound

Sound waves are longitudinal waves or waves that have the same direction of vibration as their direction of travel. The sound waves are generated by a sound source, such as the vibrating diaphragm of a stereo speaker. The sound source creates vibrations in the surrounding medium. As the source continues to vibrate the medium, the vibrations propagate away from the source at the speed of sound, thus forming the sound wave. At a fixed distance from the source, the pressure, velocity, and

displacement of the medium vary with time.

Sound waves are often simplified to a description in terms of sinusoidal plane waves, which are characterized by frequency, amplitude or sound pressure and speed of sound. The speed of sound is the distance traveled per unit time by a sound wave as it propagates through an elastic medium. In this application, the medium considered is air. At 20 °C (68 °F), the speed of sound in air is about 343 metres per second. However, this value is is not only dependent on the ambient temperature, but also varies depending on the medium through which the sound wave propagates.

2.5.2 Acoustic waves

The acoustic wave equation governs the propagation of acoustic waves through a material medium. The form of the equation is a second order partial differential equation. The equation describes the evolution of acoustic pressure p or particle velocity u as a function of position x and time t.

- **In one dimension**

  The acoustic wave equation for sound pressure in one dimension is given by Equation 2.7.

$$\frac{\partial^2 p}{\partial x^2} - \frac{1}{c}\frac{\partial^2 p}{\partial t^2} = 0 \tag{2.7}$$

  Where p is sound pressure in Pa, x is particle displacement in m, c is speed of sound in m/s and t is time in s.

  The wave equation for particle velocity has the same shape and is given by Equation 2.8 where u is particle velocity in m/s.

$$\frac{\partial^2 u}{\partial x^2} - \frac{1}{c}\frac{\partial^2 u}{\partial t^2} = 0 \tag{2.8}$$

  Equation 2.8 describes acoustic waves in only one space dimension x, because

the only other independent variable is the time t.

- **In two dimension**

  Equation 2.7 and Equation 2.8 provide mathematical equations which govern vibrations in one dimension. This subsection talks about the two dimension analog namely, the motion of an elastic membrane such as a drum head that is stretched and then fixed along its edge, membrane in a microphone or the surface of an object.

  1. **Rectangular membrane**

     An acoustic membrane is a thin layer that vibrates and is used to produce or transfer sound, such as a drum, microphone, or loudspeaker.

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2}\right) \tag{2.9}$$

     Where

$$c^2 = T/\rho$$

     In equation 2.9, T represents the tension per unit length and ρ is the mass of the undeflected membrane per unit area. The equation describes a model for obtaining the displacement u(x, y, t) of a point (x, y) on the vibrating membrane from rest (u=0) at time t.

  2. **Circular membrane**

     Circular membranes are important parts of drums, pumps, microphones, telephones, and other devices. Whenever a circular membrane is plane and its material is elastic, but offers no resistance to bending, its vibrations are modeled by the two-dimensional wave equation in polar coordinates.

$$\frac{\partial^2 u}{\partial t^2} = c^2\left(\frac{\partial^2 u}{\partial r^2} + \frac{1}{r}\frac{\partial u}{\partial r} + \frac{1}{r^2}\frac{\partial^2 u}{\partial \theta^2}\right) \tag{2.10}$$

Equation 2.10 provides a mathematical model for vibrations on a circular membrane.

2.6 Estimating structural properties of an object

When a sinusoidal excitation signal is used to induce motion in the scene objects, we can leverage the sinusoidal, time varying signal for geometric inference. Sinusoidal excitation signals will cause oscillatory vibrations on an object's surface. Traditional vibration analysis is used in structural engineering and civil engineering to estimate material properties of objects, find defects in composite materials. Typically, these involve the use of contact sensors or expensive laser vibrometers [18, 19, 20], which limit sampling to only a small number of discrete points on an object's surface.

By recovering sound from the video, it is possible to obtain a spatial measurement of the audio signal at many points on the object as opposed to a single point [47]. These spatial measurements can be used to recover the vibration modes of an object. The authors of [21] show that an object's modes of vibration are closely and predictably related to its material properties and investigate how this connection can be used to learn about the material properties of an object by analyzing its vibrations in video. Their method employs image magnification [48] and uses spatial phase variations of the complex steerable pyramid [49] to represent small local motions in video.

The authors of Davis et al.[48] state that vibration models can be useful for structural analysis. General deformations of an object can be expressed as superposition of the object's vibration modes. Vibration modes are characterized by motion where all parts of an object vibrate with the same temporal frequency, the modal frequency, and with a fixed phase relation between different parts of the object. They obtain the modal frequencies by looking for peaks in the spectra of the recovered location motion signals. At one of these peaks, they get a Fourier coefficient for every spatial location in the image. Using these Fourier coefficients, they find the vibration mode shape with amplitude corresponding to the amount of motion and phase corresponding to

fixed phase relation between points.

In modal analysis, a solid object is modeled as a system of point masses connected by springs and dampers [22]. Intuitively, rigid objects are approximated with stiff springs, highly damped objects are approximated with stiff dampers, and dense objects are approximated with heavy masses. The differential equation of motion for this system is given by:

$$m\frac{\partial^2 x}{\partial t^2} + c\frac{\partial x}{\partial t} + kx = Cam(x,t) \tag{2.11}$$

where m is mass, c is the damping coefficient, k is the spring constant and Cam(x, t) is the 1 dimensional projection of the excitation signal into the camera frame. Using Equation 2.11, observations of the motions induced by sound waves can provide information about the unknown mass, damping coefficient and the spring constant of the object. We show that is is possible to leverage the high temporal sample rate of the DVS camera to extract vibrations more easily and perform similar analysis to estimate the object's structural properties.

2.7 Summary

In this chapter, we provided a literature review of existing depth estimation methods and introduced a new acoustic-optical approach to estimate depth of scene objects using neuromorphic cameras. A brief introduction to the neuromorphic cameras, their working and applications were discussed along with the pin hole camera model for image formation. We provided equations which relate the 3D world coordinates to the 2D pixel coordinates. Finally, we discussed the acoustic wave propagation and their equations. In the subsequent chapters, we show how the topics and equations introduced in this chapter are used in the methodology proposed by this thesis.

CHAPTER 3: NEUROMORPHIC CAMERAS

This chapter provides an overview of the novel bio-inspired technology of neuromorphic cameras, or event cameras [13] used in this thesis. In contrast to standard frame based cameras, event cameras are asynchronous sensors that capture images in a completely different way. Event cameras show strong potential, when integrated with new event-driven computer vision algorithms to overcome some of the limitations of standard frame based cameras.

The chapter begins by discussing the limitations and problems caused by the design principles of conventional cameras and goes on to introduce the event based vision sensors and the motivation behind their invention. We then provide a comparison between event cameras and standard cameras and go on to describe the principle and working of different neuromorphic vision sensors.

3.1 Limitations of standard frame based cameras

Computer vision applications that involve motion of scene objects or motion of the camera itself such as in robotics or autonomous driving, the algorithms employed for standard frame based camera face several challenges and limitations, especially the response to rapid motion [50, 51] with low latency, its ability to handle extreme lighting variation [52], and loss of information between frames [53].

Standard frame-rates cannot cope with rapid motion. Many real-world, real-time vision applications require frames from cameras at a much higher rate than standard frame rates i.e. 25-60 Hz to cope with dynamics in the world. Algorithms that operate with these normal frame rates run into problems of vast motion displacement between frames or diminishing image quality due to motion blur. As shown in Figure 3.1 (a),

(a) (b)

Figure 3.1: (a) motion blur caused by rapid camera motion destroys all the detailed texture in the scene; (b) low dynamic range under extreme lighting variation, causing low contrast in the areas around bright or dark regions. Image courtesy [54]

motion blur results in the loss of detailed texture in the scene which will degrade the performance of these computer vision algorithm significantly.

Standard cameras often suffer from low dynamic range. A standard CC or CMOS camera generates video by synchronously opening its shutter to expose all pixels, or a line of pixels in the rolling shutter case, to capture frames. These sensors have relatively low dynamic range, around 60 dB and are therefore unable to sense over a high dynamic range such as in scenes that contain large intensity differences, as shown in Figure 3.1 (b). This results in a loss of visibility in an area around the bright sun and in the dark regions. Like motion blur, the low dynamic range in standard cameras will have a negative effect on computer vision algorithms.

When dealing with applications such as motion tracking [50, 51], the standard frame based cameras lose information about moving objects in the scene and in between frames. Due to the fixed frame rate of these sensors, it will not capture information about the moving object of interest. Even within each image, the same irrelevant background objects are repeatedly recorded, generating excessive unhelpful data. Thus with these sensors, static background objects end up being over sampling while motion of objects are under-sampled which leads to a loss of important scene

information.

The limitations of standard frame based cameras discussed above are of particular relevance to this thesis where we observe subtle motion on the surface of scene objects induced by sound waves. These motion occur in the order of micrometers and occur quickly in the order of microseconds. Capturing such subtle motion would not be possible with standard frame based cameras with low frame rates.

3.2 Neuromorphic Silicon Retina

Researchers in neuromorphic engineering have been trying to replicate the success of neural network architectures and functions to create electrical and electronic systems with the same efficient style of sensing and computation. Within this research branch, neuromorphic vision systems specifically aim to mimic the biological retina and subsequent vision processing. In biology, the vertebrate retina, which is a thin sheet of tissue lining the inner surface of the eye, converts raw light into electrical pulses (known as spikes) in proportion to the relative change in light intensity over time or space. The spike signals are transmitted to the brain along the optic nerve to be interpreted as visual images and to stimulate high level perception and reaction [54].

The cells in our eye report back to the brain when they detect a change in the scene. If there is no change then the cells do not report anything. The more an object moves, the more your eye and brain sample it. This process allows human vision to collect all the information it needs, without wasting time and energy reprocessing images of the unchanging parts of the scene. By only recording what changes, the eye and brain can gather useful information from things changing at up to 1000 times a second.

Motivated by this deep understanding of how visual information is encoded by the biological input sensor and transmitted to the brain, researchers in neuromorphic engineering have developed a new type of visual sensors, generally called as neuromorphic silicon retinas [55]. In 1991, Mahowald and Mead [56] successfully reproduced

the first three of the biological retina's five layers in silicon, and demonstrated the same output signals observed in real retinas in real-time. Since Mahowald and Mead's [56, 55] pioneering work, a variety of neuromorphic vision devices have been developed such as the Visio1 chip designed by Zaghloul and Boahen [57] which reproduced all five layers of the retina.

3.3 Event based vision sensors

Through historical reviews of the event-based vision sensors, research in general can be found in [13, 58, 59, 60, 61, 62]. Up to the early 2000s, neuromorphic vision research was mostly aimed at creating complete neuromorphic systems mimicking their biological counterparts as precisely as possible. Since then, many different types of silicon retina have been developed which can be used in conjunction with conventional processors and be used in practical applications as alternative vision devices. Inspired by biological vision, they extract information from scenes in various forms to reduce redundancy and latency and increase dynamic range. The event cameras originate from the bio-inspired silicon retina research in neuromorphic engineering [56, 55, 63] whose goal is to emulate some superior properties of biological vision.

Event cameras are bio-inspired vision sensors that work radically differently from conventional cameras. Event cameras measure changes in intensity asynchronously at the time they occur, rather than synchronous full image frames, as illustrated in Figure 3.2. This results in a stream of events, which encode the time, location, and polarity of brightness changes.

These cameras sample light based on the scene dynamics, rather than on a clock that has no relation to the viewed scene which gives the even camera the following advantages: very high temporal resolution and low latency (both in the order of microseconds), very high dynamic range (140 dB vs. 60 dB of standard cameras), and low power consumption. Hence, event cameras show a lot of potential for robotics and wearable applications, that present high speed and high dynamic range in the

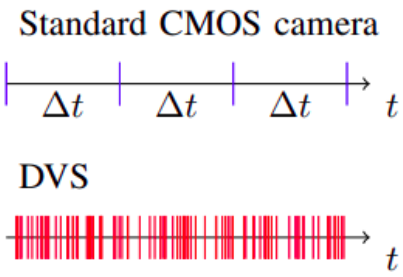


Figure 3.2: A standard CMOS camera sends images at a fixed frame rate (blue). An event camera instead sends spike events at the time they occur (red). Each event corresponds to a local, pixel-level change of brightness.

scene and which are challenging scenarios for standard cameras. Although event cameras have become commercially available only since 2008 [64], the recent body of literature [65] on these new sensors as well as the recent plans for mass production claimed by companies, such as Samsung [45] and Prophese [66], highlight that there is a big commercial interest in exploiting these novel vision sensors for mobile robotic, augmented and virtual reality (AR/VR), and video game applications.

3.4 Event camera vs. standard camera

In order to understand how event cameras work and appreciate how they can be beneficial for real time computer vision applications, it is important to look at the differences between event cameras and standard cameras, which is illustrated in Figure 3.3. Standard cameras record scenes at fixed time intervals i.e. global or rolling shutter, and output a sequence of image frames. For instance, as shown in Figure 3.3, when a fixed standard camera looks at the spinning disc with a black dot shown on the left, we get a sequence of frames as illustrated in the upper spatial-temporal graph on the right. The plot illustrates and visualizes some of the main properties of the standard video frames:

- The blind time intervals between frames, the sensor keeps sending redundant data even when the disc stays still: no new information produced
- They suffer from motion blur when the disc spins too fast, which is illustrated

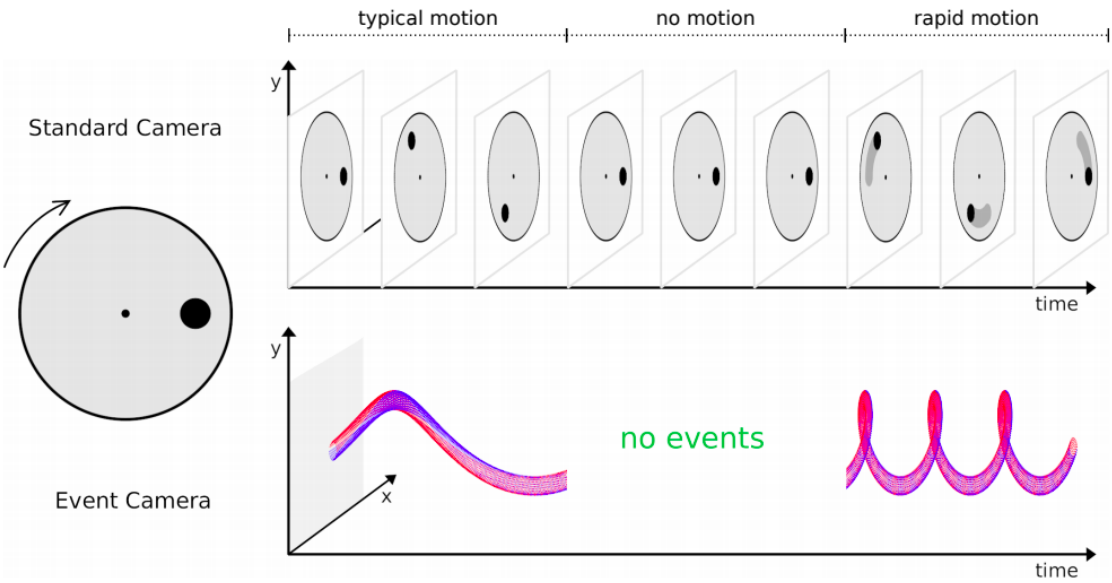


Figure 3.3: Visualization of the output of event camera vs standard camera looking at a rotating dot: in contrast to a sequence of video frames from a standard camera shown in the upper graph, a stream of events from an event camera, plotted in the lower graph, offers no redundant data output (only informative pixels or no events at all), no motion blur and high dynamic range. Red and blue dots represent positive and negative events respectively. Image courtesy [54]

by the grey tails along the trajectory of the block dot

In contrast, event cameras do not output a sequence of video frames like a standard camera, but a stream of asynchronous events, each with a pixel location, polarity and microsecond-precise timestamp, indicating when individual pixels record log intensity changes of a pre-set threshold size. Positive and negative changes produce positive or ON events and negative or OFF events respectively. By encoding only image changes, the bandwidth needed to transmit, process and store a stream of events is much lower than that for standard video, removing the redundancy in continually repeated image values.

If we observe the same spinning disc with an event camera, we get the stream of events illustrated in the lower spatial-temporal graph on the right of Figure 3.3. Red and blue dots represent positive (ON) and negative (OFF) events respectively. This graph also visualizes some of the main properties of the event stream - in particular the

almost continuous response to very rapid motion and the way that the output data-rate depends on scene motion. Hence, event cameras offer the potential to overcome the limitations of real-world computer vision applications such as low frame rate, high latency, low dynamic range, and high power consumption which conventional imaging sensors suffer from.

3.5 Principle of operation

This section discusses the working principles and the advantages and disadvantages of the 3 main types of sensors used in event cameras:

- Dynamic vision sensor (DVS)
- ATIS sensor
- Dynamic and Active pixel Vision Sensor(DAVIS)

We will discuss the operating principles and architecture of these three sensors in detail in the following subsections.

3.5.1 Dynamic vision sensor (DVS)

The Dynamic vision sensor (DVS) was the first commercialized event camera from iniLabs [67] based on the research paper of Lichtsteiner *et al.* [64]. It has a silicon retina design where the continuous-time photo-receptor was coupled to a readout circuit that was reset each time the pixel was sampled.

The DVS has a 128x128 resolution, 120 dB dynamic range and 15 μs latency, and communicates with a host computer using USB 2.0. It outputs a stream of events, each consisting of a pixel location u and v, a polarity bit p indicating either a positive or negative change in log intensity, and a timestamp t in microseconds.

Each pixel of the event camera consists of three hardware components as shown as an abstracted pixel schematic in Figure 3.4 (a): a logarithmic photoreceptor, a differencing circuit and two comparators.

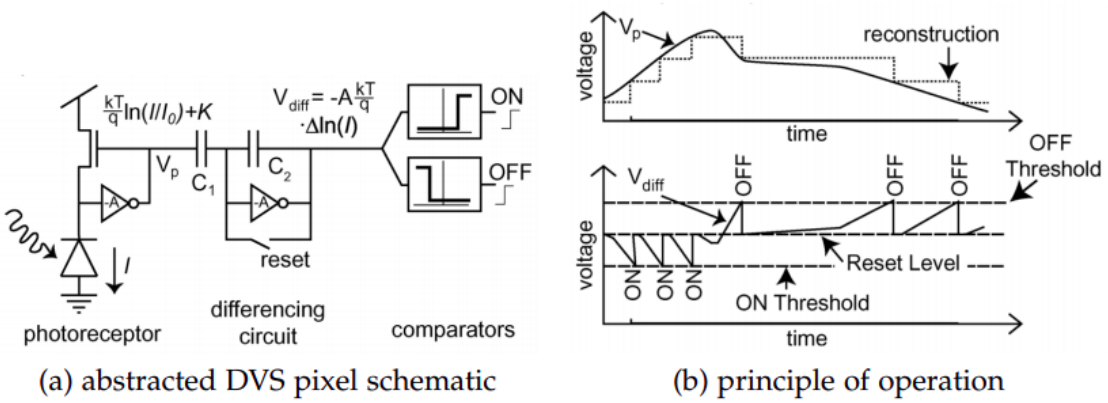


Figure 3.4: . DVS pixel architecture and the principle of operation: (a) each pixel of the DVS sensor consists of three parts: a logarithmic photoreceptor, a differencing circuit, and two comparators (b) the output signal from the photoreceptor and the principle of operation governed by the differencing circuit and comparators are plotted in the upper and lower graphs respectively. Figures courtesy of Lichtsteiner et al. [64]

The photoreceptor continuously outputs a voltage signal which encodes the incoming intensity. This signal is logarithmically as plotted in the upper graph in Figure 3.4 (b). The signal is then monitored by the differencing circuit and the two comparators for changes compared to a log intensity value recorded when the last event was emitted by the pixel. Once a change in log intensity which exceeds either a pre-set ON or OFF event threshold is detected, one of the comparators generates an ON or OFF event accordingly, and the newly generated event then resets and causes the pixel to memorize a new log intensity value as plotted in the lower graph in Figure 3.4 (b). The event threshold can be adjusted by the sensor's bias settings. The bias settings of the DVS is described in detail in Section 3.6. Each DVS pixel repeats this process continuously and independently, yielding a stream of asynchronous events which encode relative changes in pixel illumination.

3.5.2 Asynchronous Time-based Image Sensor (ATIS)

Posch et al. [53] developed the Asynchronous Time-based Image Sensor (ATIS) camera which provides absolute intensity values along with events. This was realized by combining the DVS temporal contrast pixel with a new time-based intensity mea-

surement pixel. Every scene intensity change causes three consecutive events: the first event is the same as the one from a DVS pixel, and the other two encode an absolute grey scale value in the inter-event time interval.

The main advantages of the ATIS sensors compared to the DVS sensors are their higher resolution (304×240), higher dynamic range (143 dB) and lower latency (3 μs). However, the ATIS has the disadvantage that pixels are at least double the area of DVS pixels. Additionally, in dark scenes the time between the two intensity events can be long and the readout of intensity can be interrupted by new events.

3.5.3 Dynamic and Active pixel Vision Sensor(DAVIS)

To over come the drawbacks of the ATIS sensor, Brandli et al. [15] designed the Dynamic and Active pixel Vision Sensor (DAVIS) which interleaves event data with conventional intensity frames by combining the conventional Active Pixel Sensor (APS) [68] in the same pixel with DVS. The advantage of DAVIS over ATIS is a much smaller pixel size since the photodiode is shared and the readout circuit only adds about 5% to the DVS pixel area. Intensity (APS) frames can be triggered at a constant frame rate. This circuitry is shown in Figure 3.5.

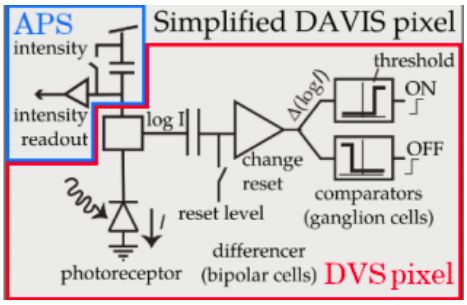


Figure 3.5: . Simplified circuit diagram of the DAVIS pixel (DVS pixel in red, APS pixel in blue). Figure courtesy of Gallego *et al.* [13]

As shown in Figure 3.5, the DAVIS pixel architecture comprises of an event based dynamic vision sensor (DVS) and a frame-based active pixel sensor (APS) in the same pixel array, sharing the same photodiode in each pixel. In this thesis, we use the DAVIS 346 [69] camera manufactured by *iniVation*. This camera has a DVS and

a gray frame resolution of 346×260 pixels, a DVS dynamic range of 120 dB and 20 μs latency, and communicates with a host computer using USB 3.0. It has bandwidth of 12 MEvents/second.

3.6 Event generation model

As discussed in the previous sections, event cameras have independent pixels that respond to changes in their log photocurrent $L \doteq log(I)$ or brightness. More explicitly, in a noise-free scenario, an event $e_k \doteq (x_k, t_k, p_k)$ is triggered at pixel $x_k \doteq (xk, yk)^T$ and at time t_k as soon as the brightness increment since the last event at the pixel, i.e:

$$\Delta L(x_k, t_k) \doteq L(x_k, t_k) - L(x_k, t_k \Delta tk) \tag{3.1}$$

reaches a temporal contrast threshold $\pm C$, as shown in Figure 3.4 (b), where $C > 0$, Δt_k is the is the time elapsed since the last event at the same pixel, and the polarity $p_k \in [-1, +1]$ is the sign of the brightness change.

3.7 Address Event Representation

The sparse and asynchronous nature of spikes (or events) in the neuromorphic design principle has called for new communication protocols which can easily distinguish which pixel has fired a specific spike or event. One such new communication strategy is the Address Event Representation (AER) [70, 56, 71], which is now the defacto standard protocol used in most neuromorphic hardware including most event cameras.

Events are transmitted from the pixel array to periphery and then out of the camera using a shared digital output bus, using the AER readout. As shown in Figure 3.6 (a), simple AER protocol-based neuromorphic systems use two control signals *req* and *ack* to synchronize the data transfer through the data bus between the sender and the receiver based on a four-phase handshake.

As illustrated in Figure 3.6 (b), the sender asserts a *req* signal once it has written

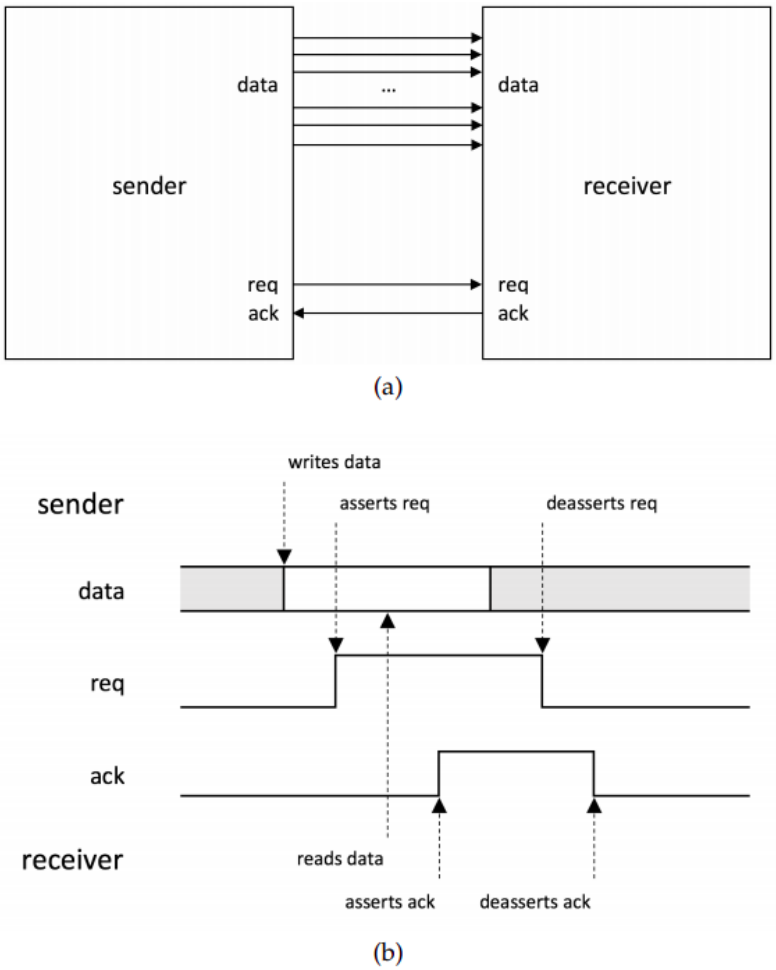


Figure 3.6: . The AER protocol: (a) the sender and the receiver of a simple AER system communicate each other through two control signals (*req* and *ack*) and a data bus; (b) a simple AER communication sequence logic diagram based on a four-phase handshake. Figure courtesy [54]

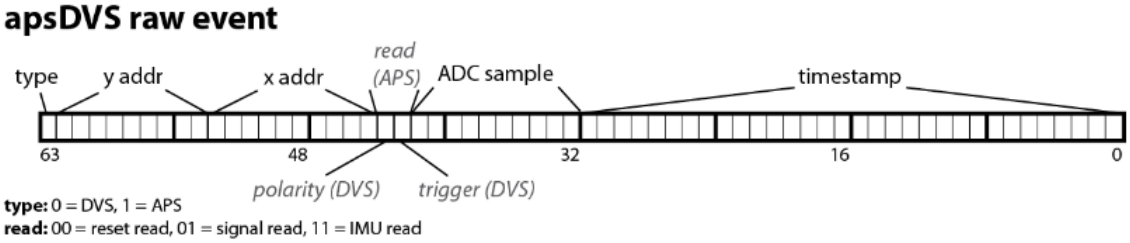
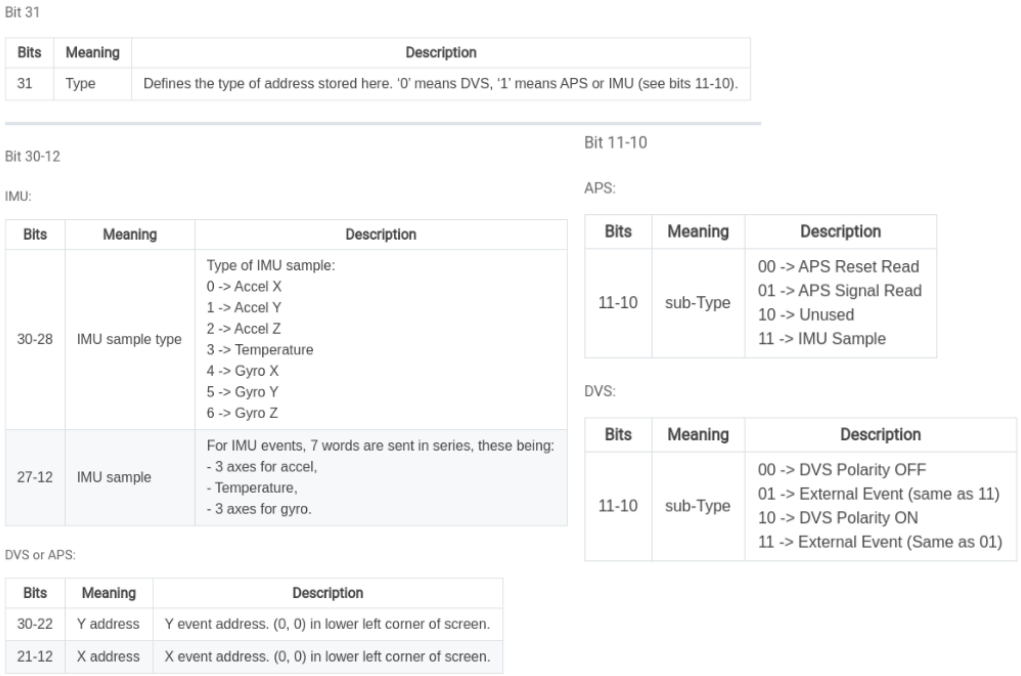


Figure 3.7: . The AER data packet for the DAVIS 346. Figure courtesy [72]

data onto the multi line data bus to notify the receiver, and at this point, the data on the bus can be considered valid. The receiver then reads the data and confirms that by asserting an *ack* signal which yields the subsequent sequential de-assertions



Bit 31

Bits	Meaning	Description
31	Type	Defines the type of address stored here. '0' means DVS, '1' means APS or IMU (see bits 11-10).

Bit 30-12

IMU:

Bits	Meaning	Description
30-28	IMU sample type	Type of IMU sample: 0 -> Accel X 1 -> Accel Y 2 -> Accel Z 3 -> Temperature 4 -> Gyro X 5 -> Gyro Y 6 -> Gyro Z
27-12	IMU sample	For IMU events, 7 words are sent in series, these being: - 3 axes for accel, - Temperature, - 3 axes for gyro.

DVS or APS:

Bits	Meaning	Description
30-22	Y address	Y event address. (0, 0) in lower left corner of screen.
21-12	X address	X event address. (0, 0) in lower left corner of screen.

Bit 11-10

APS:

Bits	Meaning	Description
11-10	sub-Type	00 -> APS Reset Read 01 -> APS Signal Read 10 -> Unused 11 -> IMU Sample

DVS:

Bits	Meaning	Description
11-10	sub-Type	00 -> DVS Polarity OFF 01 -> External Event (same as 11) 10 -> DVS Polarity ON 11 -> External Event (Same as 01)

Figure 3.8: Description of each bit stored in the DAVIS 346 AER data packet. Figure courtesy [72]

of the *req* and *ack* signals by the sender and receiver respectively, and goes back to wait for the next transaction.

The actual AER data bus width and protocol vary depending on specific hardware. For instance, the DAVIS 346 camera used in this thesis uses an AER bus whose address is 32 bit wide and the timestamp also 32 bit, for a total of 8 bytes per event, as shown in Figure 3.7. The DAVIS camera family stores polarity (luminosity change) events, IMU (Inertial Measurement Unit) samples and pixel intensity values (both APS reset and signal read) according to the following scheme shown in Figure 3.8.

3.8 DAVIS Biases

This section explains the on-chip parameters or biases of the Dynamic Vision Sensors including the DAVIS 346 camera used in this thesis. Analogue electronic circuits are often parameterised by currents or voltages which are held steady during operation and define the operating characteristics of the event camera. These currents and

Name of bias on DAVIS346	Brief description
PrBp	Photoreceptor bias - Controls the amplifier in the first stage and limits the speed with which the output of the first stage can respond to changes. This bias controls the pixel bandwidth. If the pixel bandwidth is high then it will detect faster oscillations of illumination; however, it will also respond to higher frequency electronic noise
PrSFBp	Source follower - This bias dictates the speed of the circuit whose job is to pass the signal from the first stage through to the second stage whilst reducing coupling from the second stage back to the first stage. If this bias is set sufficiently high then it should have no effect on performance. However, if this bias is low then it can limit the bandwidth of the pixel.
DiffBn	This bias controls the amplifier in the second stage and completely determines the speed at which the second stage adjusts to a change in the light-related signal.
OnBn	Threshold for On events
OffBn	Threshold for Off events
PixInvBn	Pull down for passive load inverters in digital AER pixel circuitry
RefrBp	Controls the refractory period
AEPuYBp	Pull up on request from Y arbiter
AEPuXBp	Pull up on request from X arbiter
AEPdBn	Pull down on chip request
ApsCasEpc	Cascade separating APS and DVS parts of pixel
DiffCasBnc	Cascades in differential comparator
ApsROSFBn	Source follower for column-parallel APS readout
LcolTimeoutBn	Timeout after a row event
apsOverflowLevel	APS overflow level

Figure 3.9: DAVIS 346 bias parameters [72]

voltages are called biases.

Setting these bias values is a crucial step when using an event camera. These bias settings control attributes such as sensitivity to brightness, number of On/OFF events generated, thresholds for pixel events etc. These biases can be dynamically adjusted via the camera's USB interface. Figure 3.9 lists the biases that apply to DAVIS 346 camera and describes each setting briefly.

The sensors contain digitally programmable bias generators which can produce currents that can vary over many orders of magnitude (from μA down to fA). These currents then produce voltages which can be distributed across a chip to bias many circuits at once.

3.9 Summary

In this chapter, we discussed in detail the motivation of the event camera along with its architecture and principle of operation. As shown in this chapter, the properties of event cameras offer the potential to overcome the limitations of real-time, real-world computer vision applications. The event camera is gradually becoming more widely known by researchers in computer vision, robotics and related fields and even seeing an increase in commercially available event cameras. We summarise and compare the characteristics of the notable event cameras described in this chapter in Figure 3.10. As shown in Figure 3.10, event camera technology has been improved significantly in terms of spatial resolution, pixel size, sensitivity, latency and power consumption, and we can expect many more innovations in this area in the near future. Manufacturers offer neuromorphic sensors with different resolutions, sensitivity and latencies suitable for a variety of applications.

	Supplier	iniVation			Prophesee				Samsung		
	Camera model	DVS128	DAVIS240	DAVIS346	ATIS	Gen3 CD	Gen3 ATIS	Gen 4 CD	DVS-Gen2	DVS-Gen3	DVS-Gen4
Sensor specifications	Year, Reference	2008 [2]	2014 [4]	2017	2011 [3]	2017 [66]	2017 [66]	2020 [67]	2017 [5]	2018 [68]	2020 [39]
	Resolution (pixels)	128 × 128	240 × 180	346 × 260	304 × 240	640 × 480	480 × 360	1280 × 720	640 × 480	640 × 480	1280 × 960
	Latency (µs)	12μs @ 1klux	12μs @ 1klux	20	3	40 - 200	40 - 200	20 - 150	65 - 410	50	150
	Dynamic range (dB)	120	120	120	143	> 120	> 120	> 124	90	90	100
	Min. contrast sensitivity (%)	17	11	14.3 - 22.5	13	12	12	11	9	15	20
	Power consumption (mW)	23	5 - 14	10 - 170	50 - 175	36 - 95	25 - 87	32 - 73	27 - 50	40	130
	Chip size (mm2)	6.3 × 6	5 × 5	8 × 6	9.9 × 8.2	9.6 × 7.2	9.6 × 7.2	6.22 × 3.5	8 × 5.8	8 × 5.8	8.4 × 7.6
	Pixel size (µm2)	40 × 40	18.5 × 18.5	18.5 × 18.5	30 × 30	15 × 15	20 × 20	4.86 × 4.86	9 × 9	9 × 9	4.95 × 4.95
	Fill factor (%)	8.1	22	22	20	25	20	> 77	11	12	22
	Supply voltage (V)	3.3	1.8 & 3.3	1.8 & 3.3	1.8 & 3.3	1.8	1.8	1.1 & 2.5	1.2 & 2.8	1.2 & 2.8	
	Stationary noise (ev/pix/s) at 25C	0.05	0.1	0.1	-	0.1	0.1	0.1	0.03	0.03	
	CMOS technology (nm)	350 2P4M	180 1P6M MIM	180 1P6M MIM	180 1P6M	180 1P6M CIS	180 1P6M CIS	90 BI CIS	90 1P5M BSI	90	65/28
	Grayscale output	no	yes	yes	yes	no	yes	no	no	no	no
	Grayscale dynamic range (dB)	NA	55	56.7	130	NA	> 100	NA	NA	NA	NA
	Max. frame rate (fps)	NA	35	40	NA	NA	NA	NA	NA	NA	NA
Camera	Max. Bandwidth (Meps)	1	12	12	-	66	66	1066	300	600	1200
	Interface	USB 2	USB 2	USB 3		USB 3	USB 3	USB 3	USB 2	USB 3	USB 3
	IMU output	no	1 kHz	1 kHz	no	1 kHz	1 kHz	no	no	1 kHz	no

Figure 3.10: . Comparison of common commercial event based cameras. Figure courtesy of Gallego *et al.* [13]

We introduced the DAVIS 346 camera which is used in this thesis and discussed

the attributes that control its operation such as the communication protocol used between the camera and the computer, the representation of the data generated by the camera and the programmable biases which control the operating characteristics of the event camera such as sensitivity of the sensor to light, number of events generated, thresholds for the sensors etc. These bias parameters need to be tuned appropriately to make the event camera sensitive to changes in scene illumination in order to capture the subtle vibrations generated by the acoustic excitation signal striking the surface of an object.

CHAPTER 4: METHODOLOGY

This chapter discusses the methodology used in this thesis to estimate the depth of an object using acoustic waves and neuromorphihc cameras. Our system consists of the following components:

- Acoustic wave propagation
- Camera calibration
- Geometric setup for the optical system

The first component involves emitting an acoustic wave from a speaker which will induce motion on the surface of an object. Two types of sound waves are discussed in this thesis: impulse wave and a sinusoidal wave. The next component is calibrating the camera and estimating the camera parameters including the focal length of the lens used. The final component, the primary contribution of this thesis is detecting and analyzing the vibrations caused by the acoustic excitation energy on the scene objects and setting up of a geometry for our optical system to detect these vibrations. Using this geometry, we derive an equation to calculate the depth.

4.1 Acoustic wave propagation

Sound is a pressure wave which is created by a vibrating object. These vibrations displace the particles in the surrounding medium (typical air) in an oscillatory motion, thus transporting energy through the medium. Since the particles move in parallel direction to the wave movement, sound waves are referred to as a longitudinal waves. The result of longitudinal waves is the creation of compressions and rarefactions within the air.

Let us consider sound generated by a speaker. A speaker produces sound waves by oscillating a cone, causing vibrations of air molecules. In Figure 4.1, a speaker vibrates at a constant frequency and amplitude, which produces vibrations in the surrounding air molecules. As the speaker oscillates back and forth, it results in compressing and expanding the surrounding air, creating slightly higher and lower local pressures. These compressions or high-pressure regions and rarefactions or low-pressure regions move out as longitudinal pressure waves having the same frequency as the speaker.

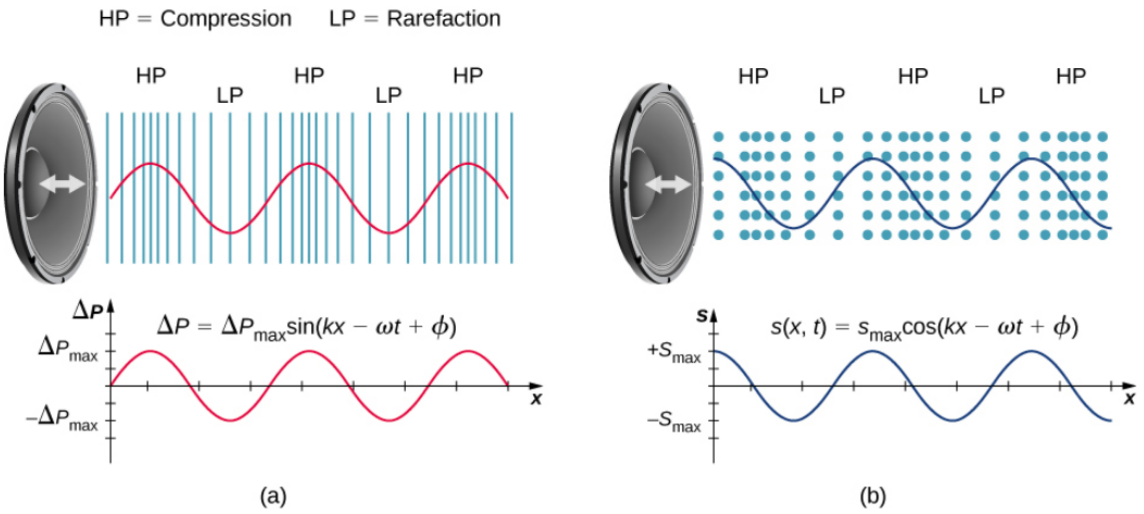


Figure 4.1: A vibrating cone of a speaker, moving in the positive x-direction, compressing and expanding the air in front of it. (a) The red graph shows the gauge pressure of the air versus the distance from the speaker. Pressures vary only slightly from atmospheric pressure for ordinary sounds. (b) Sound waves can also be modeled using the displacement of the air molecules. The blue graph shows the displacement of the air molecules versus the position from the speaker. Figure courtesy [73]

Figure 4.1(a) shows these compressions and rarefactions, and also shows a graph of gauge pressure versus distance from a speaker. As the speaker moves in the positive x-direction, it pushes air molecules, displacing them from their equilibrium positions. As the speaker moves in the negative x-direction, the air molecules move back toward their equilibrium positions due to a restoring force. The air molecules oscillate in simple harmonic motion about their equilibrium positions, as shown in Figure 4.1

(b). Since sound waves in air are longitudinal, in Figure 4.1, the wave propagates in the positive x-direction and the molecules oscillate parallel to the direction in which the wave propagates.

We can model sound waves emitted from a speaker as spherical waves from a point source. If a wave spreads out form the source in all directions, it is a three-dimensional wave. If the energy spreads uniformly in all directions in an isotropic medium (same in all directions), the wave is said to be a spherical wave. As the wave moves outward, the energy it carries is spread over a larger area.

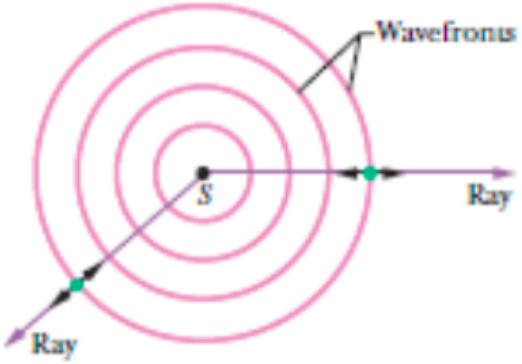


Figure 4.2: A sound wave travels from point source S

In Figure 4.2, a sound wave travels from point source S through a three dimensional medium. The wavefronts from spheres centred on S and the rays are radial to S. The short, double headed arrows indicate that elements of the medium oscillate parallel to the rays.

As described in Figure 4.1, the vibrations caused by the speaker will push the air molecules, displacing them from their equilibrium positions. If we place an object in the path of the sound waves, the displaced air molecules will cause a displacement on the surface of the object. We use this property in our thesis to induce motion in scene objects using a speaker emitting sound wave.

In this thesis, we use two types of sound waves:

- Impulse wave: An impulse-like audio wave form shown in Figure 4.3.

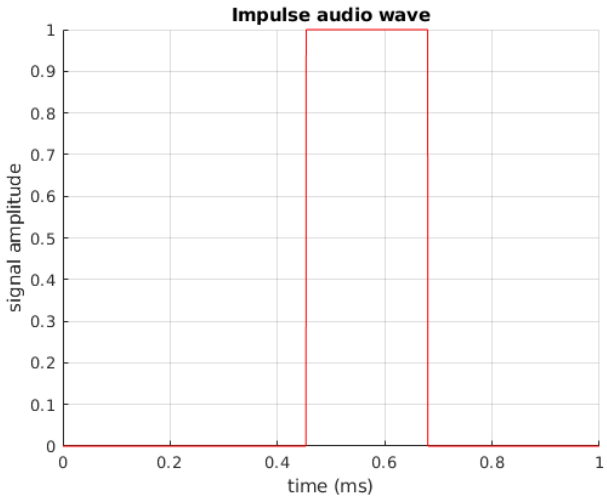


Figure 4.3: Impulse-like audio signal

Impulsive reconstruction requires only detection of motion in scene pixels. This approach does not require detection of the magnitude, phase, or frequency of the wave and depends only on the wave propagation speed, c. Using the geometrical setup described in Section 4.2, the one-way travel time, Δt or the time taken for the 3D wave to propagate from the emitter and induce a motion on observed scene locations is calculated using the AER data by detecting the time and positions at which the speaker first moves and when events are generated on the object.

- Sinusoidal wave: The second type of excitation signal used is a sinusoidal wave like the one shown in Figure 4.4. Using a sinusoidal excitation signal not only allows us to compute depth using the previously described time of flight approach but also creates similar time varying vibrations on the scene object. These vibrations can be analyzed to estimate certain structural properties of the object, which will be described in Section 5.4.

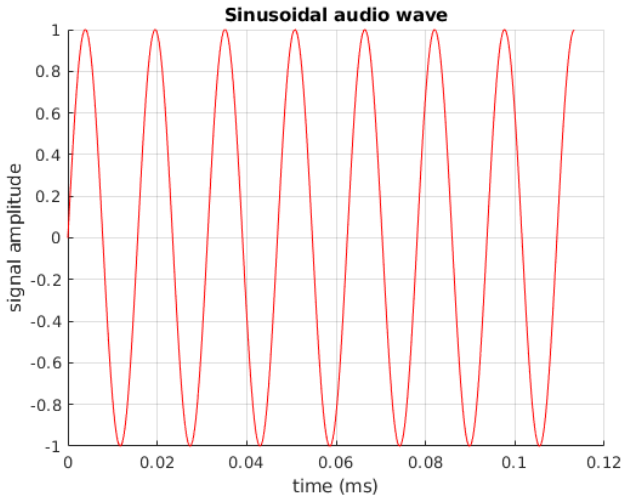


Figure 4.4: Sinusoidal audio signal

4.2 Camera calibration

Camera calibration is the process of estimating intrinsic and extrinsic parameters of a camera. Each camera in a visual system will have a unique set of camera parameters which must be determined through calibration in order to form an accurate system model. Using this model, we can develop a mathematical understanding about the relationship between points in a 3D scene and their projection onto the image plane, and how image data is represented in various frames of reference.

4.2.1 Intrinsic parameters

As discussed in Section 2.3, the pinhole camera model is used to find the 2D image pixel location or projection in the image frame of a 3D point (X, Y, Z) in the coordinate system whose origin is at the optical center of the camera. This project is expressed in terms of a set of parameters that are intrinsic to the camera. These intrinsic parameters are typically expressed in a matrix for given by:

$$K = \begin{bmatrix} \frac{F}{s_X} & 0 & c_x \\ 0 & \frac{F}{s_y} & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4.1}$$

Where (f_x, f_y) are the focal length in units pixels, F is the focal length typically expressed in mm. (p_x, p_y) is the size of a pixel in the camera sensor and is expressed in nm. This value is provided by the camera manufacturer as part of the camera of the camera specifications. (c_x, c_y) is optical center or the principal point in pixels.

Matrix K is referred to as the *camera matrix* or *intrinsic matrix* whose values we determine using through camera calibration described in Section 4.3.3.

The 2D pixel location of the 3D point is calculated using the expression:

$$\begin{bmatrix} x \\ y \\ 0 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{4.2}$$

4.2.2 Extrinsic parameters

In the perspective projection method described in the previous section, we assumed that the 3D points are expressed in the camera coordinate system, centered at the optical center or the principal point. In practice, they can be expressed in any known 3D coordinate system, which is referred as the world coordinate system. The world coordinate system can be centered about any point in the coordinate space.

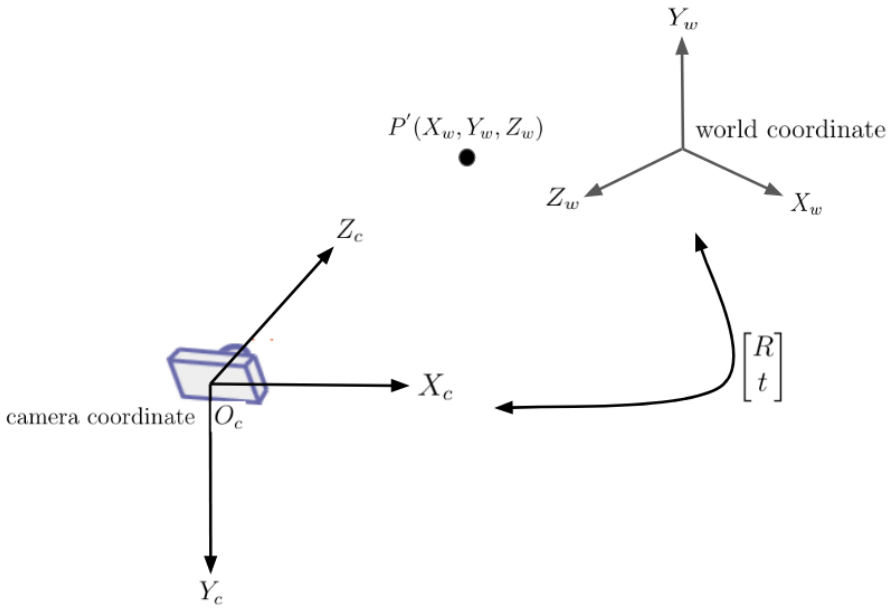


Figure 4.5: Transformation from 3D world coordinates to 3D camera coordinates

The camera's extrinsic parameters define the location and orientation of the camera with respect to this known world frame and are used to transform a 3D point in the world coordinate system to a 3D point in the camera coordinate system using the following relation:

$$P_c = R(P^{'} - t) \tag{4.3}$$

where $P^{'} = [X_w, Y_w, Z_w]^t$ is a 3D point in the world frame and $P_c = [X_c, Y_c, Z_c]^t$ is a 3D point in the camera coordinate system.

Matrix R is a 3×3 matrix known as the rotation matrix. The rotation matrix brings the corresponding axes of the two frames into alignment i.e., onto each other.

Vector t is 3×1 vector known as the translation vector between the relative positions of the origins of the two reference frames.

Equation 4.3 can be written in a homogeneous form as:

$$\begin{bmatrix} X_c \\ Y_c \\ Z_c \\ 1 \end{bmatrix} = \begin{bmatrix} R & t^T \\ O & 1 \end{bmatrix} * \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{4.4}$$

where $O = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}$ is a 1×3 zero vector.

The 2D projection of the 3D point $P^{'}$ can be calculated by combining Equation 4.3 and Equation 4.2 to get the expression:

$$P_{image} = K[R|t]P^{'} \tag{4.5}$$

or more explicitly:

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} r_{11} & r_{12} & r_{13} & t_x \\ r_{21} & r_{22} & r_{23} & t_y \\ r_{31} & r_{32} & r_{33} & t_z \end{bmatrix} \begin{bmatrix} X_w \\ Y_w \\ Z_w \\ 1 \end{bmatrix} \tag{4.6}$$

Note that if we allow the world frame to coincide with the focal point of the camera, as it did in Section 4.3.1, then then R is a 3×3 identity matrix and $t = \begin{bmatrix} 0 & 0 & 0 \end{bmatrix}^T$, then Equation 4.6 simplifies to Equation 4.2.

In Section 4.4, we will describe how these extrinsic parameters are used to calculate the ground truth value for depth in our experiments.

4.2.3 Calibration procedure

The projection matrix $P = K[R|t]$ now has 10 free parameters which we can determine through calibration. The first 4 of these parameters are the intrinsic parameters c_x, c_y, f_x, and f_y . The remaining 6 are the translation components t_x, t_y, t_z, and the 3 angles of rotation specified in R. These parameters are estimated by establishing a

set of known correspondences between world and image points.

The process of calibrating a camera to estimate these parameters is well defined in literature [74, 75]. We start by taking multiple images of a 2D pattern with known size and structure at different distances, angles and orientations. The most commonly used pattern is the checkerboard pattern like the one shown in Figure 4.6 with the dimension of each square known accurately.

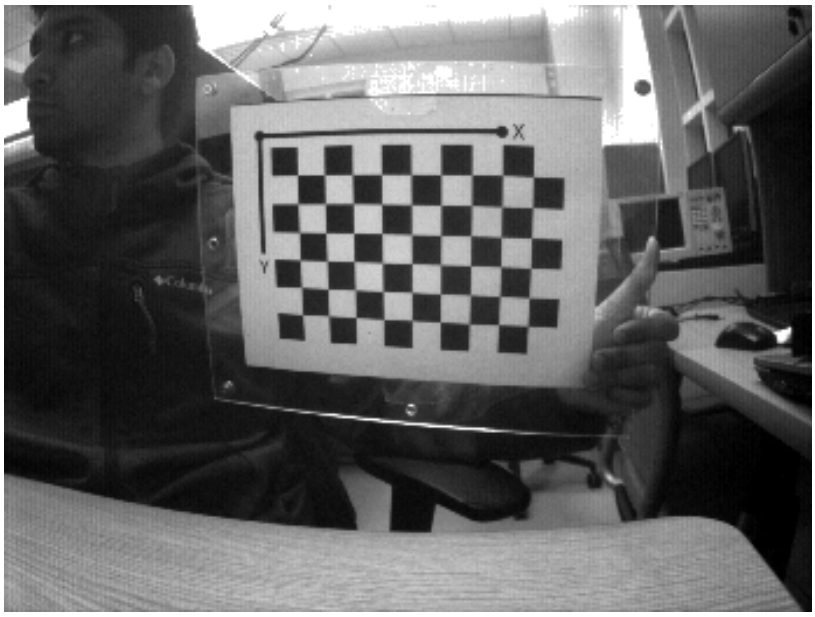


Figure 4.6: A checkerboard pattern used for camera calibration.

The pixel locations of the corners of the checkerboard (each square) are first detected. Correspondence is then computed between the physical coordinates of the pattern features (corners) and those extracted from images, resulting in a set of measurements [74, 75] that can be used to estimate the unknown parameters of the camera model, i.e. those found in the matrix K. This is done by selecting the top-left corner of the checkerboard as the origin of the world coordinate system and with x and y axis as shown in Figure 4.6. This ensures that all the detected corner points lie on the same plane and the Z component of this points is 0 in the world coordinate system.

The detected corner points and the origin for the world coordinate system is shown in Figure 4.7

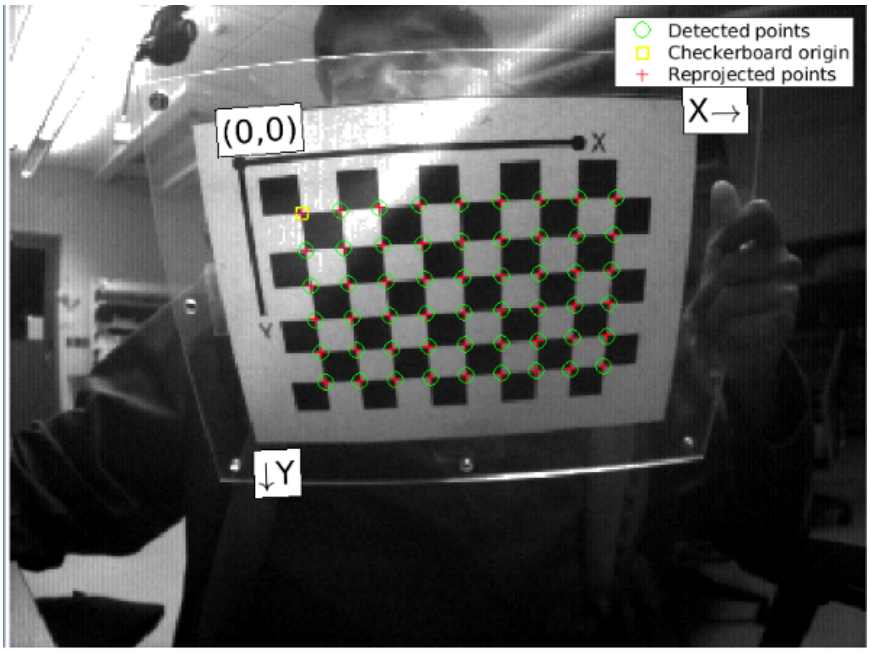


Figure 4.7: Detected corner points and origin of the world coordinate system on a checkerboard pattern.

4.2.4 Event camera calibration

Although the sensors of event cameras are fundamentally different from conventional imaging sensors, they use the same optics to which traditional perspective projection apply, as described in Section 4.2.1 and Section 4.2.2. Therefore, a calibration procedure is required for estimating the intrinsic parameters of event cameras.

When calibrating a DVS sensor, we would need to use a non static calibration pattern as the event camera only responds to scene changes. We need to either move the camera or the calibration chart, or have an active pattern such as blinking LEDs as described in [76].

However, as described in Section 3.5.3, the DAVIS sensor interleaves event data with standard intensity frames by combining the conventional Active Pixel Sensor(APS) [68] in the same pixel with DVS. Therefore, since the DAVIS 346 camera [69] is used in this thesis, the calibration procedure using a static pattern described in Section 4.2.3 can be used to calibrate the DAVIS camera using the camera's grey frames.

4.3 Geometric setup for the optical system

This section describes the geometric setup used for the optical system and how the equations for calculating depth are set up. We set up our acoustic-optical system as a triangulation problem.

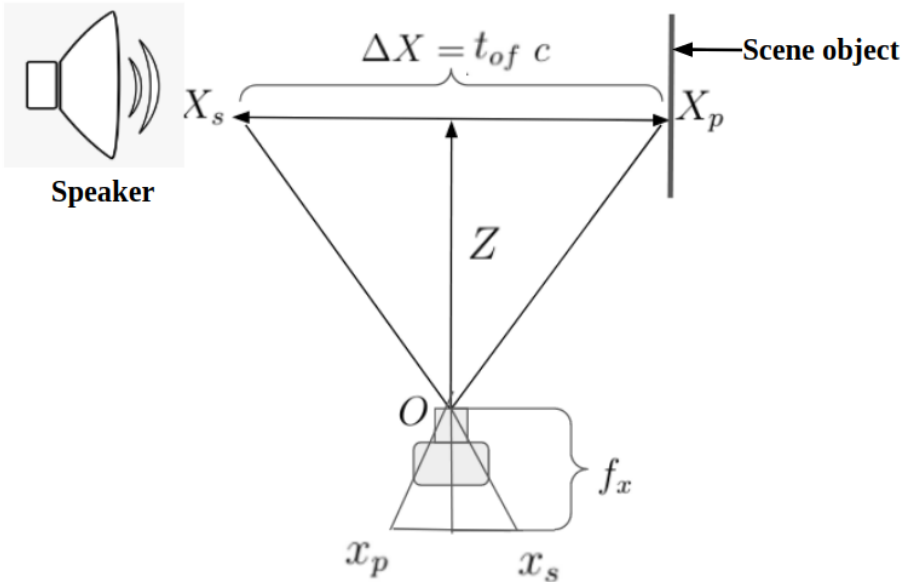


Figure 4.8: Principle of the triangulation and time of flight

Consider a speaker facing an object which is in the same plane and placed in front of the speaker. The camera is placed such that it is perpendicular to the line joining the speaker and the object. Figure 4.8 shows the top view of this geometry.

The speaker is placed at a distance ΔX from an object. The camera is placed at a distance Z as shown in Figure 4.8. If X_s and X_p are the world coordinates of the speaker and object respectively, then x_s and x_p are the projections of these points onto the image coordinate system.

The relationship between the world coordinates and the image coordinates is given by:

$$\begin{aligned} x_s &= f_x \frac{X_s}{Z} \\ x_p &= f_x \frac{X_p}{Z} \end{aligned} \tag{4.7}$$

Subtracting the two equations we get:

$$x_p - x_s = f_x \frac{X_p - X_s}{Z} \tag{4.8}$$

Equation 4.8 can be rearranged to give us the equation to calculate the depth, Z.

$$Z = f_x \frac{\Delta X}{x_p - x_s} \tag{4.9}$$

Equation 4.9 gives the value of depth in (mm) with respect to the camera's optical center. In the following steps, we discuss how each term on the right hand side of Equation 4.9 is calculated.

a. **Finding x_s and x_p**

The 2D points x_s and x_p are the pixel locations of the speaker and object in the image coordinate system. These points are detected using image processing algorithms on the grey frame from the DAVIS 346 camera. We apply a simple color threshold to segment the speaker and object locations and obtain their region of interest (ROI). We then look for events generated within these ROIs to find the location and time instances when the speaker first emitted sound waves and when motion was first induced on the surface of the object. The column locations where the events generated were first generated in the speaker and object ROIs are used as x_s and x_p.

b. **Finding f_x**

f_x is the focal length of the camera's lens with units in pixels, which we esti-

mate using the calibration procedure described in Section 4.2, specifically from Equation 4.2.

c. **Finding ΔX**

ΔX, or the distance between X_p and X_s, due to the geometry of our optical system, can be calculated simply as the distance between the x-coordinates of the two world points since the speaker and object are in the same Y-Z plane. Hence the values of $Y_p - Y_s$ and $Z_p - Z_s$ will be zero hence $\Delta X = X_p - X_s$.

To calculate this distance $X_p - X_s$, we leverage information about the speed of a sound wave and the time stamps of the events generated in the speaker and the object ROIs. Using the time stamps, we can estimate the time of flight or the time taken for the speaker to travel from the speaker to hit the surface of the object.

When the speaker emits a sound wave at time instance t_s, it will strike the object's surface and induce motion at some time instance t_p. The time of flight, t_{of} is the time taken for the sound wave to travel from the speaker to surface of the object and is calculated using Equation 4.10:

$$t_{of} = t_p - t_s \tag{4.10}$$

The distance ΔX can now be calculated by substituting t_{of} and the speed of sound in air which approximated to 343 m/s, in Equation 4.11.

$$\Delta X = (t_{of})(c) \tag{4.11}$$

With the values obtained from Steps **a**, **b** and **c** discussed above, we can simplify Equation 4.9 to:

$$Z = f_x \frac{(t_{of})(c)}{(x_p - x_s)} \tag{4.12}$$

4.4 Calculating ground truth

To obtain an accurate ground truth value of depth with respect to the camera's focal length, we take advantage of the information provided by the geometry of the 3D scene including the feature points with known 3D location and their projections from the world coordinate system back to the camera plane.

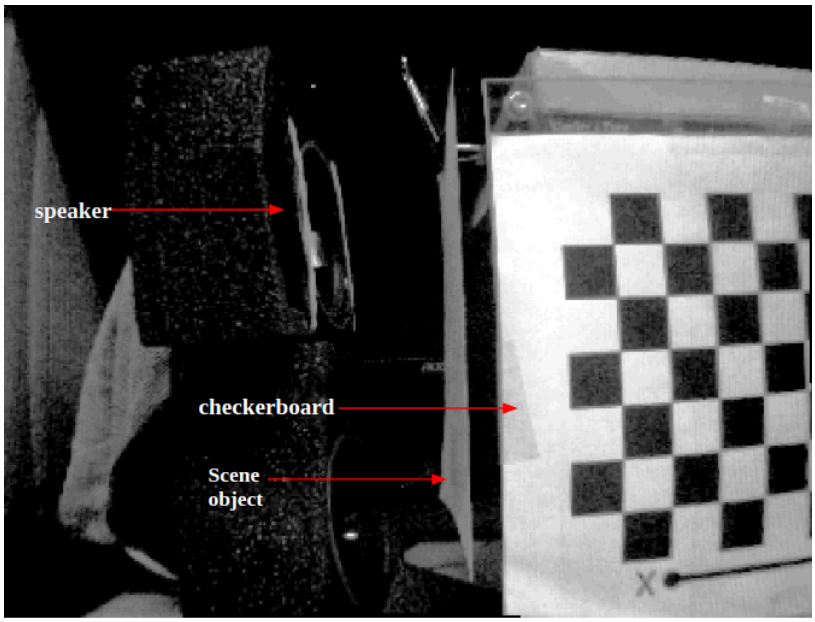


Figure 4.9: Experimental setup used to estimate ground truth of depth - as seen by the camera

To do so, we start by placing a checkerboard pattern with a known cell size behind the scene object and in the same plane as shown in Figure 4.9. The pixel locations of the corners on the checkerboard pattern are then detected. As depicted in Figure 4.10, we select the origin of the world coordinate system as the top-left corner of the checkerboard. Since the size of the checkerboard cells is known, pixel locations of the corners act as feature points with known 3D positions in the world coordinate.

Any 3D point with respect to this world coordinate system in Figure 4.10 can then be projected onto the camera coordinate system whose origin is at the the center of the image and the Z axis is aligned with the camera's optical axis. The Z coordinate of this 3D point in the camera coordinate system will be the depth value. Thus, if we

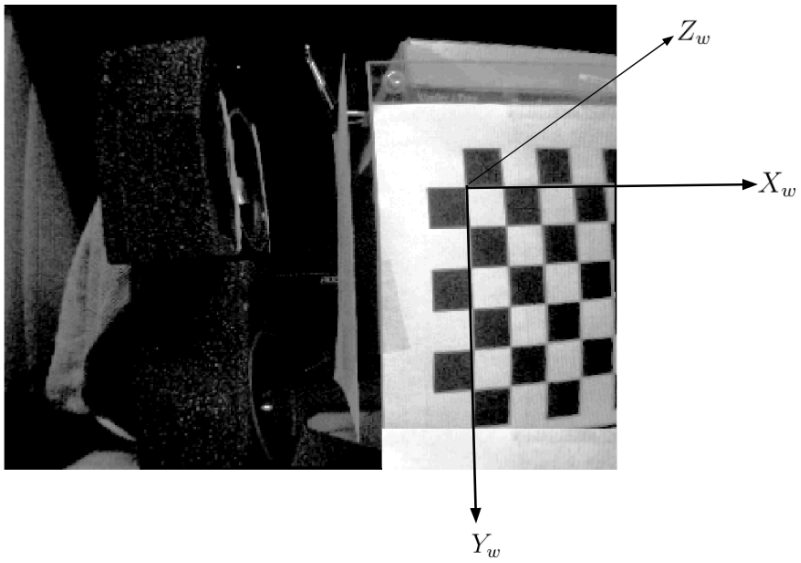


Figure 4.10: Word coordinate system with respect to the checkerboard

project the 3D world location of the object to the camera coordinate system, we can obtain the ground truth estimate for depth of the object with respect to the camera.

To transform coordinates from the world coordinate system to the camera coordinate, we use the camera's extrinsic parameters as described in Section 4.2.2 and evaluate Equation 4.4. Let $x_p = [x, y]$ be the 2D pixel location of a point on the scene object, whose depth we wish to calculate. We can then find the 3D location of x_p with respect to the world frame denoted by $[X_{pw}, Y_{pw}, Z_{pw}]$. Substituting X_{pw}, x_p, the values of the rotation matrix R and the translation vector t (which we determine from the camera calibration), in Equation 4.4, we get:

$$\begin{bmatrix} X_{pc} \\ Y_{pc} \\ Z_{pc} \\ 1 \end{bmatrix} = \begin{bmatrix} R & t^T \\ O & 1 \end{bmatrix} \begin{bmatrix} X_{pw} \\ Y_{pw} \\ Z_{pw} \\ 1 \end{bmatrix} \tag{4.13}$$

where Z_{pc} gives the ground truth value for depth.

4.5 Noise filter

All vision sensors are noisy due to the inherent shot noise in photons, transistor circuit noise, and the non-idealities present. DVS sensors also suffer from noise. The manufacturers of the DVS camera attribute three different reasons [77] why noise occurs in the stream of a DVS camera:

- Electronic noise
- Background events
- APS crosstalk

As stated by the manufacturers of the DAVIS camera [77]- "The photodiode and each of the transistors all contribute some electronic noise. In the complete absence of light there is still a small current across the photodiode - this is called the dark current; this current has a certain amount of intrinsic noise. As the light level increases, the noise in the photocurrent increases, but it does not do so exponentially. Thus, especially in low-light conditions, and in darker areas of the scene, the pixels may produce spurious events."

For background noise, the authors of [77] state that "In well-lit conditions with little electronic noise, there will nevertheless be noise events. These are all ON type, and from each pixel they arrive with a certain regularity." They attribute noise cause by APS crosstalk to a burst of excessive events when a global APS exposure is performed.

While noise can be reduced by changing certain thresholds and bias settings of the DVS sensor, it can not be eliminated completely. Having noisy pixels will impact the results of the experiments and compromise the accuracy.

Noise events, also referred to background activity (BA) noise, differ from the real activity events of a pixel such that the BA events lack temporal correlation with events in their spatial neighborhood unlike the real events that have a temporal correlation

with events from their spatial neighbors. Using this difference, the BA noise can be filtered out by detecting events generated by a pixel without the spatio-temporal correlation with the events generated by neighboring pixels and the pixel itself [78].

This filter is called a spatio-temporal correlation filter [78]. To process an event, $e_{rc}(x, y, p, t)$, the spatio-temporal filter searches e's spatial 8×8 neighborhood for events with time stamps that occur within a time interval dT with respect to t, as shown in Figure 4.11. If there exists an event with a time stamp less than a time difference dT to the processing event's time stamp, the processing event, e_{rc} has support and will pass the filter. Otherwise, the processing event will be filtered out. This principal is formulated in Equation 4.14.

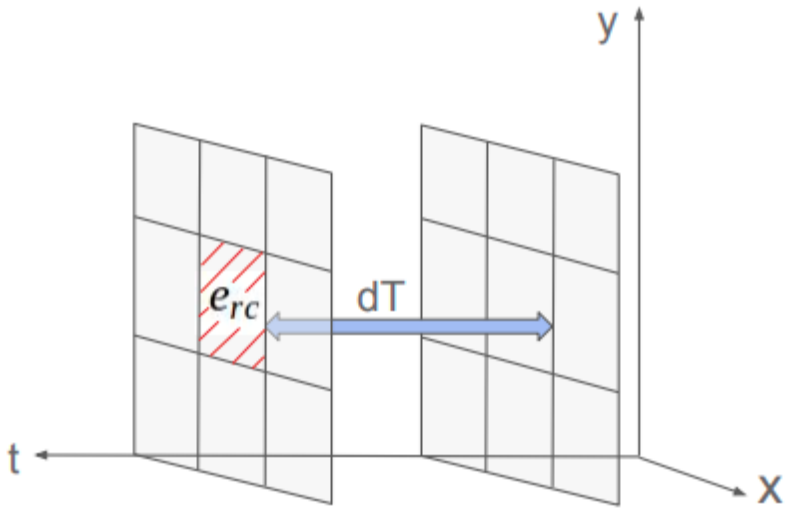


Figure 4.11: Principal of spatio-temporal noise filter. An event can pass the filter if it has correlation with its spatial neighbors within a temporal window dT.

$$e_{rc}(x, y, p, t) \neq BA \iff \exists |t - t_{ij}| < dT \quad s.t. |i - x| \leq 1 \wedge |j - y| \leq 1 \tag{4.14}$$

where e_{rc} is the processing event and t_{ij} is the time stamp of the most recent event

at $col = i$ and $row = j$, excluding the processing event.

Figure 4.12 compares an event stream before and after noise filtering.

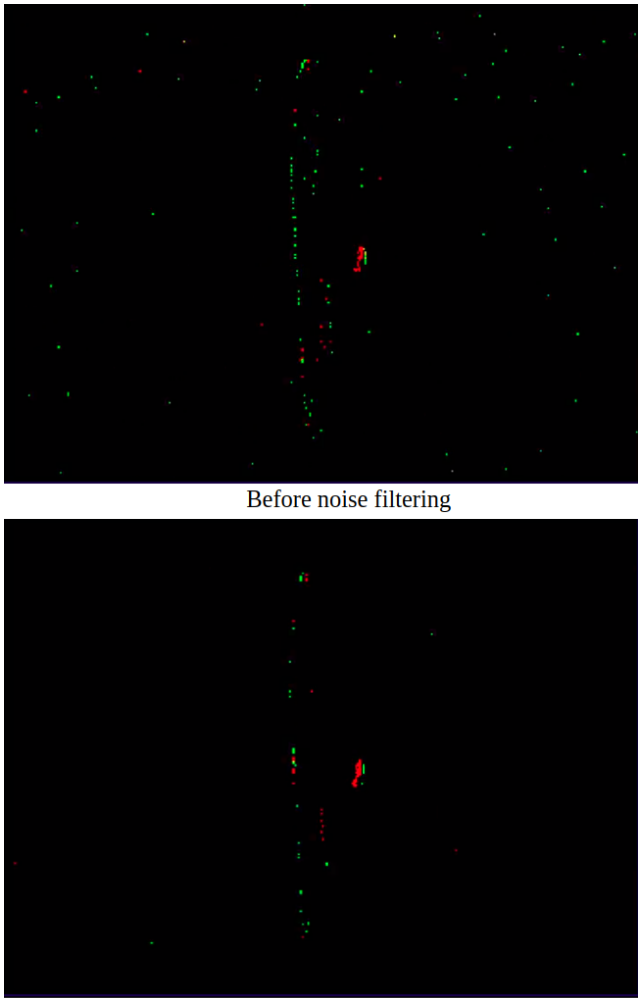


Figure 4.12: Comparison of events before and after noise filtering. The top image shows events before applying a noise filter. The image on the bottom shows the events after noise filtering

4.6 Fitting sinusoidal curve to data using discrete Fourier series

This section describes an approach to fit a sinusoidal curve to a set of noisy data points. This method is used to recover the audio signal emitted and model the movement of the speaker driver and the object in the horizontal direction.

Here, the data is an array of the column positions of the pixel corresponding to a point on the speaker or the scene object, tracked over time.

There are three steps in reconstructing the sinusoidal signal emitted: Measured data, Interpolation and Fitting. The following subsections discusses these three steps in detail.

4.6.1 Measured data

We first start by tracking an edge on the speaker and paper to get a noisy representation of their horizontal displacement. To do so, a small portion on the speaker driver was marked in white, which allows us to segment out that region using the camera's grey frame and allow us to observe events generated by the movement of this ROI. We then track the events generated by this white edge.

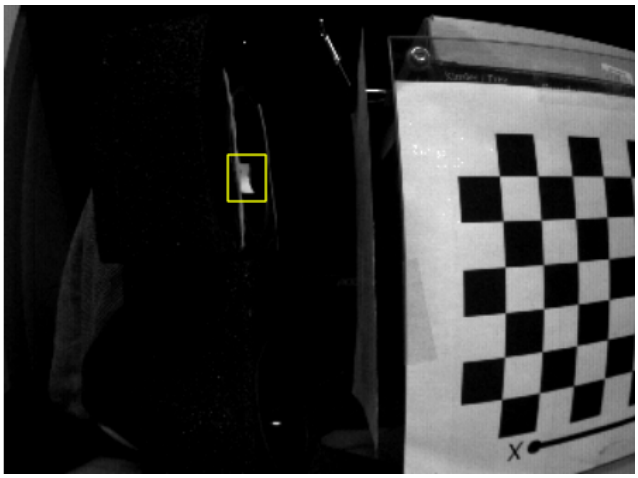


Figure 4.13: Region of interest on the speaker driver

When the speaker moves back and forth, the movement of this white region will generate events at this location. This movement is elucidated in Figure 4.14 which shows a simulation of the white region moving back and forth.

Figure 4.14: Movement of the white region on the speaker driver

In Figure 4.14, (**a**) the region represents the white ROI marked on the speaker

driver with the red edge representing the right most edge of the white marking. The movement of the white region moving to the left is shown in (**b**), (**c**) and (**d**). As the white region moves back, the surrounding speaker region will be visible in its place, which creates a black region. As the black region begins to appear, the event camera observes this change as a negative change in intensity and will generate OFF events on the right hand side of the red line and generate ON events on the left hand side of the of the red line since the white region moving back causes a positive change in intensity. Similarly, as the white region moves to the right, back to its initial rest position, the event camera will generate ON events to the right side of the red line as the white region begins to occupy the black background which is sensed as a positive change in intensity and generate OFF events to the right of this red line.

Figure 4.15: Events generated by movement of speaker driver

The events generated for the driver moving back and forth is shown in Figure 4.15. Using the events generated, it is possible to identify the edge of the white region (the red line in Figure 4.14). Keeping track of the the position of this edge on this horizontal axis over time will give us the horizontal displacement of the speaker which we use as the measured data. It is important to note that the edge must be tracked at a rate greater than $\frac{1}{2f}$, in accordance to Nyquist's theorem, where f is the frequency of the signal.

4.6.2 Interpolation

In practice, events measured are not equally spaced in time. The measured data described in the previous section will be noisy due to the missed detections of the edge, or events not being generated frequently due to lighting conditions or sparse samples of the measured data.

Hence, we perform a linear interpolation on the measured data so that the data is uniformly-spaced on the time axis before we fit a sinusoid.

4.6.3 Fitting

We can define this problem as fitting a sinsoid of known frequency to the measured data. The Fourier series fits N sinusoids which are harmonics of a finite length signal. The harmonics here are the frequencies $k\omega_0$ where $\omega_0 = \frac{2\pi}{N}$. We wish to fit a specific frequency ω_c. To fit a sinusoid with this frequency, we substitute ω_c for $k\omega_0$.

We start off with the equations for Fourier series, whose synthesis equation is given by:

$$f(t) = \frac{a_0}{2} + \sum_{k=1}^{\infty} \left(a_k \cos(k\Omega_0 t) + b_k \sin(k\Omega_0 t)\right) \tag{4.15}$$

The analysis equations are given by:

$$\begin{aligned} a_k &= \frac{2}{T} \int_T x(t) \cos(k\Omega_0 t) dt \\ b_k &= \frac{2}{T} \int_T x(t) \sin(k\Omega_0 t) dt \end{aligned} \tag{4.16}$$

We then convert these equations to their discrete form. In Equation 4.16, T indicates the period whose discrete form is N, t becomes nT and $\Omega_0 T$ will give us the discrete frequency, ω_0. Thus, the discretized form of Equation 4.16 is:

$$\begin{aligned} a_k &= \frac{2}{N}\sum_{n=0}^{N-1} x[n]\cos(k\omega_0 n) \\ b_k &= \frac{2}{N}\sum_{n=0}^{N-1} x[n]\sin(k\omega_0 n) \end{aligned} \tag{4.17}$$

If we consider L to be the number of periods, then Equation 4.17 becomes:

$$\begin{aligned} La_k &= \frac{2}{N}\sum_{n=0}^{LN-1} x[n]\cos(k\omega_0 n) \\ \implies a_k &= \frac{2}{LN}\sum_{n=0}^{LN-1} x[n]\cos(k\omega_0 n) \\ Lb_k &= \frac{2}{N}\sum_{n=0}^{LN-1} x[n]\sin(k\omega_0 n) \\ \implies b_k &= \frac{2}{LN}\sum_{n=0}^{LN-1} x[n]\sin(k\omega_0 n) \end{aligned} \tag{4.18}$$

Therefore, for a specific frequency, ω_c and L number of periods, the discretized analysis equations are:

$$\begin{aligned} a_n &= \frac{2}{LN}\sum_{n=0}^{LN-1} x[n]\cos(\omega_c n) \\ b_n &= \frac{2}{LN}\sum_{n=0}^{LN-1} x[n]\sin(\omega_c n) \end{aligned} \tag{4.19}$$

We now have the Fourier series coefficients for the form shown in Equation 4.15. The sine and cosine pairs in Equation 4.15 can be expressed as a single sinusoid with a phase offset defined by:

$$f(t) = \frac{a_0}{2} + \sum_{n=1}^{LN-1} A_n\cos(\omega_c t + \phi) \tag{4.20}$$

where A_n and ϕ are given by:

$$\begin{aligned} A_n &= \sqrt{{a_n}^2 + {b_n}^2} and \\ \phi &= arctan(\frac{b_n}{a_n}) \end{aligned} \tag{4.21}$$

Thus, we get the equation for a single sinusoid with the desired frequency, ω_c, which will give us a sinusoid curve fit to the measured data.

4.7 Summary

In this chapter, we have presented the mathematical notation conventions and geometrical foundations used throughout this thesis. We propose a geometric setup for the optical system and derive Equation 4.9 to estimate depth based on this geometry, which is the primary contribution of this thesis. We discuss how to estimate all the terms on the right hand side of Equation 4.9 using the output of the event camera and the acoustic-optical setup.

Additionally, we also design and describe an approach to estimate ground truth value for the experiments. Finally, we describe how Equation 4.19, Equation 4.20 and Equation 4.21 can be used to extract an estimate of the acoustic wave excitation from the apparent motion of the scene objects using the event camera data.

CHAPTER 5: RESULTS

This chapter describes the experimental set up used for the thesis and provides the specifications and results of all the components of the system.

5.1 Experimental setup

Figure 5.1 shows the setup used to estimate depth and an overview of the algorithm used to calculate depth is shown in Algorithm 1. The scene object used for the experiments was a sheet of paper with dimensions 190×130 mm. A speaker, emitting sound waves is placed in front of this paper to induce motion.

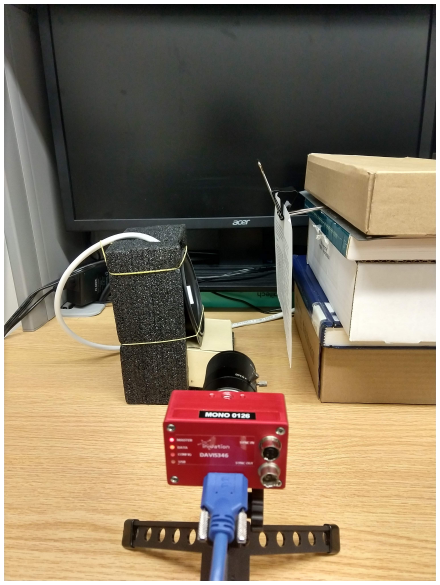


Figure 5.1: Experimental setup used

To detect the time instance (t_s) at which the sound wave was emitted, a small region on the diaphragm of the speaker is detected using the DAVIS camera's grey frames. Once this region of interest (ROI) is detected, the algorithm checks to see if any events are generated within this ROI. If an event is detected, this event's time

Algorithm 1: Algorithm used to calculate depth using time of flight

```
Load AER data;
RemoveNoise();
loadCalibration();
while event(ID) <= numofevents do
    if event(ID).x >= speakerROI.x and event(ID).x <= speakerROI.x +
     speakerROI.width and event(ID).y >= speakerROI.y and event(ID).y <=
     speakerROI.y + speakerROI.height then
        ts = event(ID).timeStamp;
        xs = event(ID).x;
    end
    if notisempty(ts) and event(ID).x >= paperrROI.x and event(ID).x <=
     paperrROI.x + paperrROI.width and event(ID).y >= paperrROI.y and
     event(ID).y <= paperrROI.y + paperrROI.height then
        tp = event(ID).timeStamp;
        xp = event(ID).x;
        toff = (tp - ts);
        X = toff * c;
        deltax = xp-xs;
        fx = DAVIScameraParams.FocalLength(1)*Px;
        Z = (fx*X)/deltax;
    end
    ID = ID+1;
end
```

stamp is saved as t_s and the location on the horizontal axis is saved as x_s. After the speaker start time is detected, the algorithm checks for events generated in the right half of the image (where the object is in the frame). When an event is detected here, the time stamp is saved as t_p and location on the horizontal axis is saved as x_p. Equation 4.9 is then evaluated using the camera's focal length (f_x) to find the depth, Z in millimeters (mm).

5.2 Camera calibration

To calibrate the camera using the method described in Section 4.3.3, MATLAB's calibration procedure [79] was used. The results of the calibration are shown below:

$$K = \begin{bmatrix} \frac{F}{s_X} & 0 & c_x \\ 0 & \frac{F}{s_y} & c_y \\ 0 & 0 & 1 \end{bmatrix} = \begin{bmatrix} 251.427 \pm 0.3129 & 0 & 171.5425 \pm 0.3070 \\ 0 & 252.6674 \pm 0.3060 & 116.6541 \pm 0.2809 \\ 0 & 0 & 1 \end{bmatrix} \tag{5.1}$$

The pixel size of the Davis346 camera is provided by the manufacturers [72] and is $18.5 \times 18.5 \mu m$. This indicates a variance of $\sim 6\ \mu m$ in the results of the calibration.

5.3 Emitted wave type

For the time of flight calculation, two types of emitted waves were used to induce displacement in the scene objects: impulse-like audio wave and a sinusoidal wave.

5.3.1 Impulse wave

Figure 5.2 shows a plot of the horizontal displacement of the speaker driver and paper over time. The column indices of the pixels where movement was first detected on the speaker and paper are shown along with their time stamps. The difference between these two times instances is used as the time-of-flight or t_{of} in Equation 4.11.

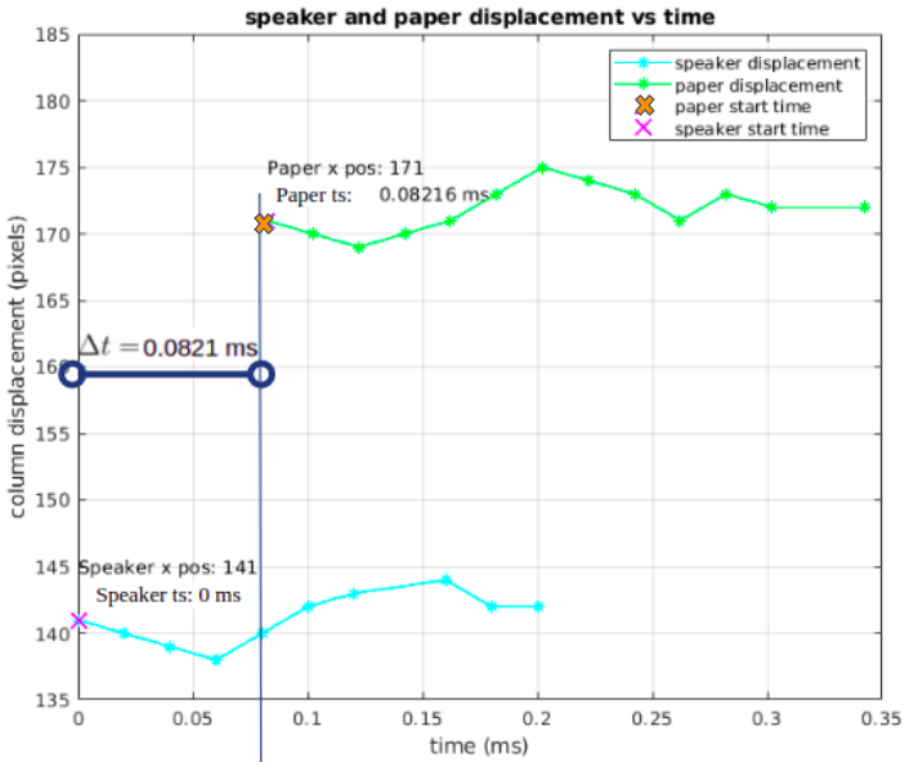


Figure 5.2: Plot of the horizontal displacement of the speaker driver and paper over time

Using the calibration values from Equation 5.1 and time-of-flight information from Figure 5.1, the depth is calculated using Equation 4.12 as shown below:

$$\begin{aligned} f_x &= 251.427 \\ x_p &= 171 \\ x_s &= 141 \\ t_{of} &= 0.0821ms \\ Z &= f_x \frac{t_{of} \times c}{x_p - x_s} \\ Z &= 235.72 \pm 5.632 \times 10^{-3} mm \end{aligned} \tag{5.2}$$

The results for impulsive depth reconstruction experiments are shown in Table 5.1.

Table 5.1: Results for depth estimation with impulse wave

	ground truth (mm)	estimated depth (mm)
1	231.3	$235.7 \pm 5.632 \times 10^{-3}$
2	375.3	$379.8 \pm 5.632 \times 10^{-3}$
3	282.2	$271.9 \pm 5.632 \times 10^{-3}$
4	248.4	$236.2 \pm 5.632 \times 10^{-3}$

All ground truth were calculated using the method described in Section 5.2.

5.3.2 Sinusoidal wave

For these experiments, a low frequency sine wave was used. A sine wave with frequency of 64 Hz and sampled at 44100 Hz was used for all experiments.

The results for depth estimation using the time of flight approach with sinusoidal signals is shown in Table 5.2

Table 5.2: Results for depth estimation with sine wave

	ground truth (mm)	estimated depth (mm)
1	312.3	$309.5 \pm 5.632 \times 10^{-3}$
2	237.2	$238.2 \pm 5.632 \times 10^{-3}$
3	303.6	$300.9 \pm 5.632 \times 10^{-3}$
4	254.7	$256.6 \pm 5.632 \times 10^{-3}$

Figure 5.3 shows the result of fitting a sinusoidal curve to the measured data - column positions of the pixel corresponding to the speaker driver, as described in Chapter 4.6.3. The fit curve, shown in red indicates we are able to retrieve the emitted audio signal.

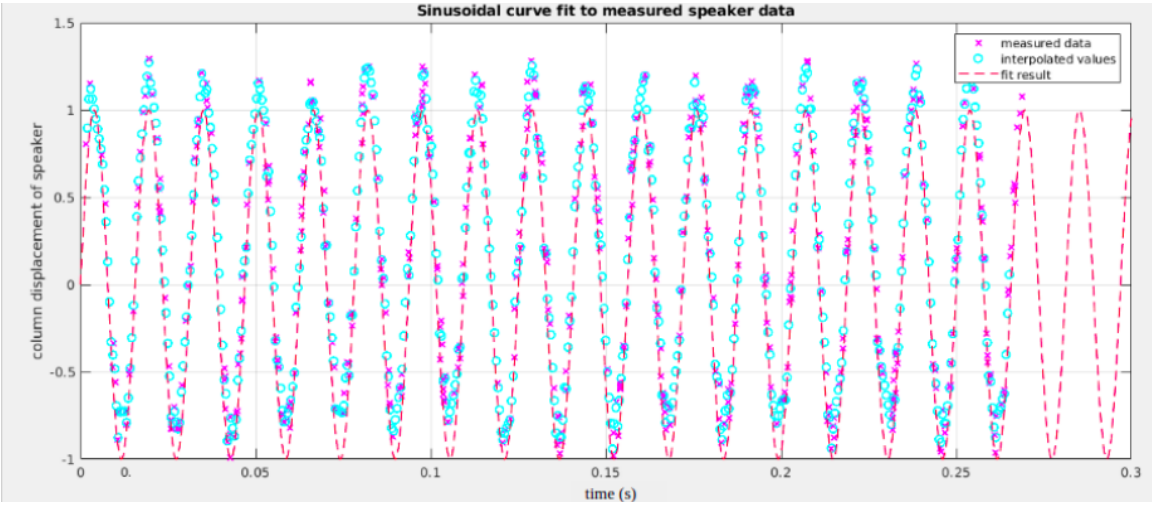


Figure 5.3: Result of fitting sinusoidal curve to measured data of speaker movement

The fitting experiment described in Chapter 4.6.3 allows us to extract the frequency of the excitation signal in Hertz. As the measured data is the horizontal displacement of the scene object in terms of pixels over time, the result of the fit sinusoid curve is also plotted as the horizontal displacement of the object in pixels (y axis in Figure 5.3), over time (x axis in Figure 5.3). Hence the amplitude of the fit sine curve is expressed in terms of pixels. For this experiment, a sinusoid with frequency 64 Hz was used as the excitation signal. From Figure 5.3, we can calculate the average peak to peak time of the fit sinusoid which is found to be 15.1 ms thus resulting in an estimated frequency of 66.2 Hz and an amplitude of 1 which indicates the displacement of the scene object in pixels in each direction of the horizontal axis. This indicates an error of 2.2 Hz in our estimate of the frequency with respect to the true frequency of the excitation signal.

5.4 Summary

In this chapter, we described the experimental setup used in this thesis and provided the experimental results obtained. We provide the calibration results and show from the calibration results that a variance of $\sim 6\ \mu m$ is applied to all measurements. With the help of Figure 5.2 and Equation 5.2 we elucidate how to use the event camera data

and the camera parameters to solve for the variables in Equation 4.12 to calculate depth. We also present the result of the fitting experiment to estimate the frequency of the sinusoidal excitation signal which estimates the frequency with an error of 2.2 Hz.

CHAPTER 6: CONCLUSIONS

In this thesis, we successfully developed a novel non-contact, monocular depth estimation method using an acoustic excitation signal and a calibrated neuromorphic camera sensor. Within the taxonomy of existing depth estimation methods, this approach can be classified as an acoustic-optical sensing mechanism - a novel and one of its kind method to estimate depth.

The methodology described in this thesis takes advantage of the high temporal resolution and high frame rates of event cameras to observe and calculate the time taken by an acoustic wave to strike an object's surface. This information is used in conjunction with the proposed geometry for the optical system to recover depth. We derive an equation to estimate depth using this geometry. The results show that the proposed method is able to evaluate the depth accurately with an error of $\pm 1cm$.

We also propose a method to extract the frequency of a sinusoidal excitation signal by observing the motion in the scene using the event data. The results indicate that the proposed method estimates the frequency of the excitation signal with an error of 2.2 Hz. This approach for reconstructing the emitted signal along with the acoustic-optical sensing mechanism proposed in this thesis show potential for estimating structural properties of the object and has use cases in fields such as vibration analyses, estimating structural proprieties, robotics etc.

REFERENCES

[1] D. Scharstein, R. Szeliski, and R. Zabih, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," in *Proceedings IEEE Workshop on Stereo and Multi-Baseline Vision (SMBV 2001)*, pp. 131–140, 2001.

[2] Z. Ma and S. Liu, "A review of 3d reconstruction techniques in civil engineering and their applications," *Advanced Engineering Informatics*, vol. 37, pp. 163 – 174, 2018.

[3] T. Lindeberg and J. Garding, "Shape from texture from a multi-scale perspective," in *1993 (4th) International Conference on Computer Vision*, pp. 683–691, May 1993.

[4] J. Malik and P. Perona, "Preattentive texture discrimination with early vision mechanisms," *J. Opt. Soc. Am. A*, vol. 7, pp. 923–932, May 1990.

[5] R. Zhang, P. sing Tsai, J. E. Cryer, and M. Shah, "Shape from shading: A survey," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 21, pp. 690–706, 1999.

[6] A. Maki, M. Watanabe, and C. Wiles, "Geotensity: Combining motion and lighting for 3d surface reconstruction," *International Journal of Computer Vision*, vol. 48, pp. 75–90, 07 2002.

[7] Z. Zhang, "Microsoft kinect sensor and its effect," *IEEE MultiMedia*, vol. 19, pp. 4–10, Feb 2012.

[8] C. O. S. Quegan, *Understanding Synthetic Aperture Radar Images*. Artech House, 1998.

[9] Y. Lu, K. Yang, and R. Duan, "A simple method for depth estimation of a sound source at known range in the deep sea," *The Journal of the Acoustical Society of America*, vol. 146, no. 6, pp. 4097–4107, 2019.

[10] R. Hansen, "Synthetic aperture sonar technology review," *Marine Technology Society Journal*, vol. 47, pp. 117–127, 09 2013.

[11] Chiang-Jung Pu, J. G. Harris, and J. C. Principe, "A neuromorphic microphone for sound localization," in *1997 IEEE International Conference on Systems, Man, and Cybernetics. Computational Cybernetics and Simulation*, vol. 2, pp. 1469–1474 vol.2, Oct 1997.

[12] M. Crocco, A. Trucco, and A. Del Bue, "Uncalibrated 3d room reconstruction from sound," 06 2016.

[13] G. Gallego, T. Delbrück, G. Orchard, C. Bartolozzi, B. Taba, A. Censi, S. Leutenegger, A. J. Davison, J. Conradt, K. Daniilidis, and D. Scaramuzza, "Event-based vision: A survey," *ArXiv*, vol. abs/1904.08405, 2019.

[14] T. Delbruck and M. Lang, "Robotic goalie with 3ms reaction time at 4% cpu load using event-based dynamic vision sensor," *Frontiers in neuroscience*, vol. 7, p. 223, 11 2013.

[15] C. Brandli, R. Berner, M. Yang, S. Liu, and T. Delbruck, "A 240×180 130 db 3 μs latency global shutter spatiotemporal vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 49, no. 10, pp. 2333–2341, 2014.

[16] A. Amir, B. Taba, D. Berg, T. Melano, J. McKinstry, C. D. Nolfo, T. Nayak, A. Andreopoulos, G. Garreau, M. Mendoza, J. Kusnitz, M. Debole, S. Esser, T. Delbruck, M. Flickner, and D. Modha, "A low power, fully event-based gesture recognition system," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 7388–7397, July 2017.

[17] Z. Ni, A. Bolopion, J. Agnus, R. Benosman, and S. Regnier, "Asynchronous event-based visual shape tracking for stable haptic feedback in microrobotics," *IEEE Transactions on Robotics*, vol. 28, no. 5, pp. 1081–1089, 2012.

[18] V. Aranchuk, A. K. Lal, C. F. Hess, and J. M. Sabatier, "Multi-beam laser Doppler vibrometer for landmine detection," *Optical Engineering*, vol. 45, no. 10, pp. 1–10, 2006.

[19] O. Büyüköztürk, R. Haupt, C. Tuakta, and J. Chen, "Remote detection of debonding in frp-strengthened concrete structures using acoustic-laser technique," in *Nondestructive Testing of Materials and Structures* (O. Güneş and Y. Akkaya, eds.), (Dordrecht), pp. 19–24, Springer Netherlands, 2013.

[20] J. Chen, R. Haupt, and O. BÃŒyÃŒkozturk, "The acoustic-laser vibrometry technique for the noncontact detection of discontinuities in fiber reinforced polymer-retrofitted concrete," *Materials Evaluation*, vol. 72, pp. 1305–1313, 10 2014.

[21] A. Davis, K. L. Bouman, J. G. Chen, M. Rubinstein, O. Buyukozturk, F. Durand, and W. T. Freeman, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, pp. 732–745, 04 2017.

[22] A. A. Shabana, *Theory of Vibration Volume 2: Discrete and Continuous Systems*. Berlin, Germany: Springer, 1991.

[23] B. Horn, "Obtaining shape from shading information," *Shape from Shading*, pp. 123–171, 08 1989.

[24] A. Hartley, R.; Zisserman, *Multiple View Geometry in Computer Vision*. Cambridge University Press: Cambridge, UK, 2003.

[25] A. Global, "Xtion pro," October 2019.

[26] Orbbec, "Orbbec astra," October 2019.

[27] T. Delbruck and P. Lichtsteiner, "Fast sensory motor control based on event-based hybrid neuromorphic-procedural system," in *2007 IEEE International Symposium on Circuits and Systems (ISCAS)*, pp. 845–848, May 2007.

[28] D. Drazen, P. Lichtsteiner, P. Hafliger, T. Delbruck, and A. Jensen, "Toward real-time particle tracking using an event-based dynamic vision sensor," *Experiments in Fluids*, vol. 51, pp. 1465–1469, 11 2011.

[29] T. Delbruck and M. Lang, "Robotic goalie with 3ms reaction time at 4% cpu load using event-based dynamic vision sensor," *Frontiers in neuroscience*, vol. 7, p. 223, 11 2013.

[30] A. Glover and C. Bartolozzi, "Event-driven ball detection and gaze fixation in clutter," *Int Conf Intell Robot Syst IROS*, pp. 2203–2208, 2016.

[31] A. Glover and C. Bartolozzi, "Robust visual tracking with a freely-moving event camera," *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 3769–3776, 2017.

[32] M. Litzenberger, A. N. Belbachir, N. Donath, G. Gritsch, H. Garn, B. Kohn, C. Posch, and S. Schraml, "Estimation of vehicle speed based on asynchronous data from a silicon retina optical sensor," *IEEE Intell. Transp. Sys. Conf.*, pp. 653–658, 2006.

[33] E. P. A. N. B. S. Schraml, M. G. B. K. C. Posch, and S. Schraml, "Spatiotemporal multiple persons tracking using dynamic vision sensor," *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pp. 35–40, 2012.

[34] G. W. S. S. M. L. A. N. B. M. Hofstatter and C. Bartolozzi, "Event-driven embodied system for feature extraction and object recognition in robotic applications," *IEEE Conf. Comput. Vis. Pattern Recog. Workshops (CVPRW)*, pp. 76–82, 2012.

[35] G. Orchard, C. Meyer, R. Etienne-Cummings, C. Posch, N. Thakor, and R. Benosman, "Hfirst: A temporal approach to object recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 37, pp. 2028–2040, Oct 2015.

[36] 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, pp. 1346–1359, July 2017.

[37] A. Sironi, M. Brambilla, N. Bourdis, X. Lagorce, and R. Benosman, "Hats: Histograms of averaged time surfaces for robust event-based object classification," pp. 1731–1740, 06 2018.

[38] J. H. Lee, T. Delbruck, M. Pfeiffer, P. K. J. Park, C. Shin, H. Ryu, and B. C. Kang, "Real-time gesture interface based on event-driven processing from stereo

silicon retinas," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, pp. 2250–2263, Dec 2014.

[39] P. Rogister, R. Benosman, S. Ieng, P. Lichtsteiner, and T. Delbruck, "Asynchronous event-based binocular stereo matching," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 23, pp. 347–353, Feb 2012.

[40] E. Piatkowska, A. N. Belbachir, and M. Gelautz, "Asynchronous stereo vision for event-driven dynamic stereo sensor using an adaptive cooperative approach," in *2013 IEEE International Conference on Computer Vision Workshops*, pp. 45–50, Dec 2013.

[41] H. Rebecq, G. Gallego, E. Mueggler, and D. Scaramuzza, "Emvs: Event-based multi-view stereo—3d reconstruction with an event camera in real-time," *International Journal of Computer Vision*, vol. 126, pp. 1394–1414, Dec 2018.

[42] Z. Xie, S. Chen, and G. Orchard, "Event-based stereo depth estimation using belief propagation," *Frontiers in Neuroscience*, vol. 11, p. 535, 2017.

[43] A. Andreopoulos, H. J. Kashyap, T. K. Nayak, A. Amir, and M. D. Flickner, "A low power, high throughput, fully event-based stereo system," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7532–7542, June 2018.

[44] G. Gallego, H. Rebecq, and D. Scaramuzza, "A unifying contrast maximization framework for event cameras, with applications to motion, depth, and optical flow estimation," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3867–3876, June 2018.

[45] B. Son, Y. Suh, S. Kim, H. Jung, J. Kim, C. Shin, K. Park, K. Lee, J. Park, J. Woo, Y. Roh, H. Lee, Y. Wang, I. Ovsiannikov, and H. Ryu, "A," in *2017 IEEE International Solid-State Circuits Conference (ISSCC)*, pp. 66–67, Feb 2017.

[46] C. Brandli, R. Berner, M. Yang, S. Liu, and T. Delbruck *IEEE Journal of Solid-State Circuits*, vol. 49, pp. 2333–2341, Oct 2014.

[47] A. Stanbridge and D. Ewins, "Modal testing using a scanning laser doppler vibrometer," *Mechanical Systems and Signal Processing*, vol. 13, 08 2002.

[48] A. Davis, M. Rubinstein, N. Wadhwa, G. Mysore, F. Durand, and W. T. Freeman, "The visual microphone: Passive recovery of sound from video," *ACM Transactions on Graphics (Proc. SIGGRAPH)*, vol. 33, no. 4, pp. 79:1–79:10, 2014.

[49] J. Portilla and E. Simoncelli, "A parametric texture model based on joint statistics of complex wavelet coefficients," *International Journal of Computer Vision*, vol. 40, 10 2000.

[50] T. Delbruck and P. Lichtsteiner, "Fast sensory motor control based on event-based hybrid neuromorphic-procedural system," in *2007 IEEE International Symposium on Circuits and Systems*, pp. 845–848, 2007.

[51] E. P. Simoncelli, W. T. Freeman, E. H. Adelson, and D. J. Heeger, "Shiftable multiscale transforms," *IEEE Transactions on Information Theory*, vol. 38, no. 2, pp. 587–607, 1992.

[52] H. Rebecq, R. Ranftl, V. Koltun, and D. Scaramuzza, "High speed and high dynamic range video with an event camera," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[53] C. Posch, D. Matolin, and R. Wohlgenannt, "A qvga 143 db dynamic range frame-free pwm image sensor with lossless pixel-level video compression and time-domain cds," *IEEE Journal of Solid-State Circuits*, vol. 46, no. 1, pp. 259–275, 2011.

[54] H. Kim, *Real-Time Visual SLAM with an Event Camera.* PhD thesis, Imperial College London - Department of Computing, September 2017. Last retrieved 2020-05-09.

[55] M. A. Mahowald and C. Mead, "The silicon retina.," *Scientific American*, vol. 264 5, pp. 76–82, 1991.

[56] M. Mahowald, *VLSI Analogs of Neuronal Visual Processing: A Synthesis of Form and Function.* PhD thesis, California Institute of Technology, May 1992. Last retrieved 2020-05-09.

[57] K. A. Zaghloul, *A Silicon Implementation of a Novel Model for Retinal Processing.* PhD thesis, University of Pennsylvania, 2001. Last retrieved 2020-05-09.

[58] T. Delbruckl, "Neuromorophic vision sensing and processing," pp. 7–14, 09 2016.

[59] T. Delbruck, "Fun with asynchronous vision sensors and processing," in *Proceedings of the 12th International Conference on Computer Vision - Volume Part I*, ECCV'12, (Berlin, Heidelberg), pp. 506–515, Springer-Verlag, 2012.

[60] T. Delbruck, B. Linares-Barranco, E. Culurciello, and C. Posch, "Activity-driven, event-based vision sensors," in *Proceedings of 2010 IEEE International Symposium on Circuits and Systems*, pp. 2426–2429, 2010.

[61] C. Posch, "Bio-inspired vision," *Journal of Instrumentation*, vol. 7, pp. C01054–C01054, jan 2012.

[62] C. Posch, T. Serrano-Gotarredona, B. Linares-Barranco, and T. Delbruck, "Retinomorphic event-based vision sensors: Bioinspired cameras with spiking output," *Proceedings of the IEEE*, vol. 102, no. 10, pp. 1470–1484, 2014.

[63] B. K, "Neuromorphic chips.," *Scientific American*, vol. 262, pp. 56–63, 2005.

[64] Lichtsteiner *et al.*, "A 128×128 120 db 15 μs latency asynchronous temporal contrast vision sensor," *IEEE Journal of Solid-State Circuits*, vol. 43, no. 2, pp. 566–576, 2008.

[65] Robotics and U. o. Z. Perception Group, *Event-based Vision Resources*, 2020 (accessed May 5, 2020).

[66] Prophesee, *Event-based cameras by Prophesee*, 2020 (accessed May 5, 2020).

[67] iniLabs, *DVS camera*, 2020 (accessed May 5, 2020).

[68] E. R. Fossum, "Cmos image sensors: electronic camera-on-a-chip," *IEEE Transactions on Electron Devices*, vol. 44, no. 10, pp. 1689–1698, 1997.

[69] inivation, *DAVIS346 camera*, 2020 (accessed May 5, 2020).

[70] J. Lazzaro, J. Wawrzynek, M. Mahowald, M. Sivilotti, and D. Gillespie, "Silicon auditory processors as computer peripherals," *IEEE Transactions on Neural Networks*, vol. 4, no. 3, pp. 523–528, 1993.

[71] M. A. Sivilotti, *Wiring Considerations in Analog VLSI Systems, with Application to Field-Programmable Networks.* PhD thesis, California Institute of Technology, 1990. Last retrieved 2020-05-09.

[72] inivation, *DAVIS346 camera AER file format*, 2020 (accessed May 10, 2020).

[73] L. Learning, *Sound Waves*, 2020 (accessed May 11, 2020).

[74] R. Tsai, "A versatile camera calibration technique for high-accuracy 3d machine vision metrology using off-the-shelf tv cameras and lenses," *IEEE Journal on Robotics and Automation*, vol. 3, no. 4, pp. 323–344, 1987.

[75] Z. Zhang, "A flexible new technique for camera calibration," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 11, pp. 1330–1334, 2000.

[76] E. Mueggler, N. Baumli, F. Fontana, and D. Scaramuzza, "Towards evasive maneuvers with quadrotors using dynamic vision sensors," pp. 1–8, 09 2015.

[77] inivation, *DVS event based camera biasing and noise*, 2020 (accessed March 28, 2020).

[78] A. Khodamoradi and R. Kastner, "O(n)-space spatiotemporal filter for reducing noise in neuromorphic vision sensors," *IEEE Transactions on Emerging Topics in Computing*, pp. 1–8, 2017.

[79] Mathworks, *Camera calibration using MATLAB*, 2020 (accessed May 5, 2020).