Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione

Algorithms for 3D Data Estimation from Single-Pixel ToF Sensors and Stereo Vision Systems

Master candidate:
Ufuk Baran KARAKAYA

Advisor:
Associate Prof. Pietro ZANUTTIGH

Company Coordinator:
Dr. Matteo PERENZONI

Master of Science in
ICT for Internet and Multimedia

Department of Information Engineering
University of Padova
2022

Abstract

Depth Map Estimation from stereo devices and time of flight range cameras has been a challenging issues in Computer Vision. Distance Estimations from single-pixel histograms of time-of-flight sensors are exploited in numerous fields. Beyond the several drawbacks such as degradation caused by strong ambient light, scattered and multi-path possibilities, most of the prediction algorithms could be applied to resolve these problems effectively. As these two different tasks are handled in connection with each other, supervised approaches are considered since they provide more robust results. These results are used to train the model to improve three-dimensional geometry information and against major difficulties such as complicated patterns and objects. These approaches are observed according to their accuracy with help of metrics and get improved their performances.

This thesis focuses on the analysis of Time-of-Flight and stereo vision systems for depth map estimation and single-pixel distance prediction. State of art algorithms are compared and implemented with additional strategies which are integrated to minimize the error ratio. The histograms which are obtained from Time of Flight Sensor Simulation are exploited as a dataset for single-pixel distance prediction and after that, NYU Dataset is selected for depth map estimation.

Sommario

La stima della mappa della profondità da dispositivi stereo e telecamere del tempo di volo sono stati problemi impegnativi in Computer Vision. Le stime della distanza da istogrammi a pixel singolo dei sensori di tempo di volo sono sfruttate in numerosi campi. Al di là dei numerosi inconvenienti tali poiché il degrado causato dalla forte luce ambientale e le possibilità sparse e multi-percorso, la maggior parte degli algoritmi di previsione potrebbe essere applicata per risolvere efficacemente questi problemi. Poiché questi due diversi compiti sono gestiti in connessione l'uno con l'altro, vengono presi in considerazione approcci supervisionati poiché forniscono risultati più solidi. Questi risultati vengono utilizzati per addestrare il modello a migliorare le informazioni sulla geometria tridimensionale e contro le maggiori difficoltà come modelli e oggetti complicati. Questi approcci vengono osservati in base alla loro accuratezza con l'aiuto di metriche e migliorano le loro prestazioni.

Questa tesi si concentra sull'analisi dei sistemi di visione del tempo di volo e stereo per la stima della mappa di profondità e la previsione della distanza a pixel singolo. Gli algoritmi all'avanguardia vengono confrontati e implementati con strategie aggiuntive integrate per ridurre al minimo il rapporto di errore. Gli istogrammi ottenuti dalla simulazione del sensore del tempo di volo vengono sfruttati come set di dati per la previsione della distanza a pixel singolo e, successivamente, viene selezionato il set di dati NYU per la stima della mappa di profondità.

Contents

1 Introduction

Analysis and recognition of the environment in which we live, for human perception, is considered extremely simple. When the accuracy of this estimation is handled how it is robust, there are several variables such as color, shadow, light, and object structures. All these parameters are exploited for human depth sense as a part of 3D geometry with high sensitivity.

In spite of Image Processing and Computer Vision branches have been focusing on how the visual systems work, this subject has not been explained in full detail. Meanwhile, many techniques, scientific approaches have been developed to recover the 3D geometry of objects in imagery. Stereo devices and time of flight sensors are examples of the most common systems for depth estimation. Stereo devices were developed with gained inspiration from the human vision system. It is based on the difference between two adjacent cameras which record the pictures to merge. After that, the time of flight technology is introduced as a method that can calculate the distance directly for every scene. In recent times, some additional devices are released such as Microsoft Kinect, they lead increasing of usage 3D data estimation. It provides us new possibilities to improve the results and increase the usage fields.

Essentially, Stereoscopic Imaging is a method for creating or enhancing the illusion that an image has depth by showing two slightly offset images separately to each eye of the viewer. These images belong to the same scene but they have some differences between each other such as visual angle and perspective. It does not provide us with more robust results for fundamental problems (e.g textureless fields). Nevertheless, it is exploited as one of the most common techniques. Especially, it handles 3-D information and high-resolution color images efficiently.

Following then, time-of-flight cameras were made available. A time of flight camera is a device that uses ToF measurement to estimate distances between the camera and objects or surroundings in order to create pictures made up of individually measured points. It determines depth information using infrared light (lasers that are invisible to the naked eye). It was inspired by bats' ability to discern distance. The sensor sends out a light signal that is intercepted by the item and returned to the target. The time it takes to bounce back is then calculated, allowing for depth mapping. We may think of laser-based scanner-less LIDAR imaging systems, motion sensing and tracking, object recognition for machine vision and autonomous driving, topography mapping, and more as ToF camera applications. As the applications of a ToF camera, we may consider laser-based scanner-less LIDAR imaging systems, motion sensing and tracking, object detection for machine vision and autonomous driving, topographic mapping, and more.

Even if they need more power consumption and they have more production cost, they solve drawback of stereo cameras such as the geometry of textureless object and occlusion problem of stereo systems.The most common disadvantages of these systems are low resolution and fail of success about background illumination, multi-path and scattered light scenarios. In addition, building ToF camera is complex due to calibration requirement.

Both of these two technologies have some drawbacks and advantages on different aspects. Single Pixel Distance Detection and Image Depth Maps need these divergent benefits to provide a sufficient result. For this reason, usage of these methods for the different distance and data estimation contribute to improve results.

In this thesis, the main technical details and principles of time of flight algorithm and stereo devices being exploited for distance and depth estimation on single pixel and images are investigated. Both of these approaches are analyzed and observed with help of hyper-parameters for the related datasets. The 2nd Chapter covers a general description for time of flight camera and stereo vision devices. The 3rd Chapter discusses more details for related techniques, after that the 4th, 5th and 6th Chapters concern these technologies separately according to related datasets (single pixel distance detection, image based depth maps for stereo and monocular devices). First, working principles are introduced after that structure of datasets and pre-operations and then the algorithms which are exploited to estimate the distance and depth map are merged with essential details of these methods.

Finally, Results of these algorithms,conclusions and possible future works are presented in the 7th Chapter.

2
Literature Review

Time of Flight Algorithms have been used to handle a range of challenges in Image Processing and Computer Vision. According to [1], it may be applied to both direct and indirect variations. Direct ToF (dToF) is a technique for determining the amount of time required for a reflection to occur. Indirect ToF (iToF) calculates distance by collecting reflected light and measuring the phase shift between emitted and reflected light. The phase shift of the received beat or balanced light is used to determine the depth data in i-ToF devices. This is typically accomplished using pixel-level photo-demodulators, and in reality requires the acquisition of three or more sub-frames, obtained sequentially, and appropriately combining modulation-demodulation schemes to reduce background and increase target reflectivity. After that, it determines the depth. I-ToF is an excellent choice for short-range, high-resolution 3D imaging, particularly when a solid foundation is nearby (bounded by pixel full well capacity).

On the other hand, [2] concentrates on the correction method for indirect Time of Flight Sensors and emphasizes the critical role of single-photon LIDAR in depth imaging. Their approach is based on the creation of a histogram using the time delays between generated light pulses and observed photon arrivals.

[3] mentions that the histogram may be used for picture segmentation. It classifies a picture based on the density of its pixels into one of many predefined types. It is employed recurrent neural networks as a recursive technique for resolving distance mistakes for image segmentation. The study generates a learnable histogram layer from the photos. This

learnable histogram approach can be adaptable to increase the accuracy of depth estimation as well.

In [4], robust posture estimation utilizing Time of Flight sensors is based on depth information. It does this by maximizing the quality of the findings by utilizing the confidence value for each pixel in the 3D range picture. It utilizes depth range sensors to determine the modulation frequency's confidence value.

For the image depth map part, [5] is about an algorithm that is based on re-projecting ToF data on the stereo camera viewpoint and up-samples the data. The spatial resolution of stereo devices is improved by combinational segmentation and filtering. The proposed method aims to merge the time of flight camera values and stereo devices values to obtain better confidence information for depth estimation.

Instead of additional resources and merged architectures in [6] is based on only time of flight sensors with filters which are used to ignore the external factor such as sunlight and reflectivities. It aims to reduce the dataset dimension without losing the crucial details for 3D data estimation. Nevertheless, the degree of data aggregation for the pose estimation process may lead to longer run-time and precision progress. To support these steps with help of monochromatic images, the algorithms which work for 3D data are applied (e.g. ICP, SIFT, KLT).

In [7] follows the data transfer technique to obtain better results in monocular depth estimation. Instead of standard stereo camera data, it is based on image pairs extracted from the stereoscopic film. This situation requires a sufficient approach for disparity extraction because standard disparity approaches are designed for stereo images. For this reason, smaller disparity ranges cannot be handled effectively. The academic study is based on an optical flow algorithm to overcome this issue.

In [8] focuses on image segmentation with increasing the depth information quality via convolutional neural networks. It is based on encoder-decoder architecture. Two different branches are dedicated to features that are based on RGB and depth images. After these decoder steps, features have been merged with each other. Comprehensive results illustrate that these fusion-based proposed techniques provide us results that are as sufficient as the results of the state-of-art method.

[9] offers a network that successfully utilizes the spatio-temporal structures of ToF frequency data. This method yields more robust findings for the performance of the join multi path elimination, denoising, and phase unwrapping method on a variety of difficult problems.

3
SET-UP FOR TIME OF FLIGHT SYSTEMS

This chapter is based on a crucial general introduction to both two systems before the explanation of single-pixel estimation and image-based depth map estimation. Even if stereo devices are commonly used in our lives, especially the ToF is a newer technology for which practical applications are starting to appear(e.g it is in the last iPhones). This part aims to provide a sufficient introduction to the systems which are used in this thesis and some technical expressions, and general concepts.

3.1 Fundamentals for ToF

The essential logic of time of flight sensors is based on distance calculation with help of illumination of the target object. The reflection which returns from the target object is analyzed. The first part of the thesis involves this sensor system and considers its technology for the implementation of the single-pixel distance.

The picture makes a point about the operation of time-of-flight devices. The distance L between target and sensor is calculated using the time T required for an electromagnetic wave to cross that distance. As a scientific calculation, the speed of light is assumed to be constant in the air ($c = 3x10^8 m/s$). Thus, the distance between target and sensor is calculated as the product of light speed and the time interval between them. The needed hardware is developed in accordance with the ideas in [10], consisting of a radiation emitter (TX) and a

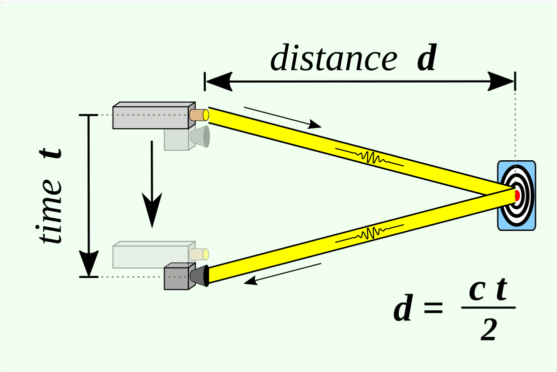


Figure 3.1

receiver (RX), which are preferably co-located. At time t = 0, the transmitter releases a light pulse that travels directly toward the scene for a distance L until it reaches the target at time t= T/2. After reaching the destination, it is reflected back to the source t=T and detected by RX. The relationship between time and distance in this case enables us to determine the distance of a single point from the Time-of-Flight sensor.

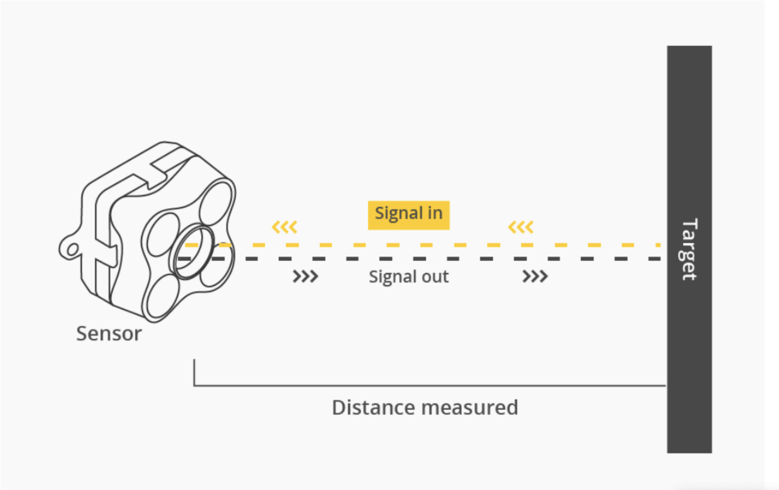


Figure 3.2: Time-of-Flight Sensor

This technology is defined as LIDAR technology, It is exploited for the creation of high-resolution depth maps in spectacular wide fields. Time-Flight sensors detect the distance between two constant points. For this reason, it cannot be adapted for dynamic systems efficiently.

Time of Flight Camera systems are a more advanced version of LIDAR sensors, with a matrix of N x M ToF sensors measuring the relevant scene. It provides the quick availability of depth maps, which is critical for real-time applications. Although there is no direct relationship between emitters and receivers owing to physical constraints, IR LEDs can be arranged in a regular pattern to simulate a single emitter in the receiver matrix's center. We can get depth information without doing any additional computations by integrating N x M receivers on a single CMOS chip, which is reasonably compact and easy to build. The primary shortcomings of this design are its low resolution in contrast to a conventional camera, poor quality across the region with depth discontinuities, and extreme sensitivity to light fluctuations. Additionally, it can be affected by other light sources in the area, such as the sun, or by an item's lack of reflectivity.

3.2 Fundamentals for Stereo Devices

Stereo systems are defined as a framework that is generated from two regular cameras(identical), which are inspired by human stereopsis. A stereo vision system is a computer vision implementation that uses stereoscopic ranging techniques to estimate a 3D model of the scene.

Stereo vision employs triangulation, a classic range technique, to compute depth from 2D images. Stereopsis is defined as binocular vision, in which our brain combines information about this three-dimensional structure gathered by our eyes from two slightly different perspectives of the same image. This idea may be extended to cameras that capture the same scene but are separated by a set distance, such as human eyes. The left camera, denoted by L, serves as a reference point, and the right camera, denoted by R, serves as a target.

3D coordinates of a point may be measured by triangulation of the correspondent points. The system starts with a standard form of two cameras (aligned and parallel), and then measures a point in the space $\mathbf{P} = [x, y, z]$. The projections of the left and right cameras are calculated $p_l = [u_l, v_l]$ and $p_r = [u_r, v_r]$ as left and right. Especially, The process of triangulation handles the determining the coordinates of P for depth information.

At that point, it is clearly understandable that the only difference in the coordinates of p_l and p_r is horizontal coordinate (u), vertical coordinates (v) will not be changed.

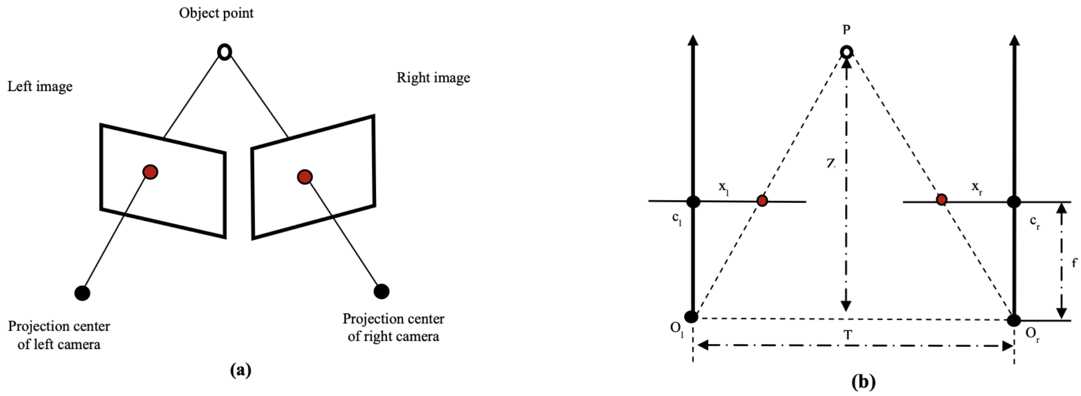
Given the geometry depicted and similar triangles properties, the following equations can be derived.

$$\frac{f}{z} = \frac{-u_l}{x} \tag{3.1}$$

$$\frac{f}{z} = \frac{u_r}{x-b} \tag{3.2}$$

$$z = \frac{bf}{u_r - u_l} = \frac{bf}{d} \tag{3.3}$$

After the some calculations are applied to Formulae 3.1 and 3.2, The expression is given above are obtained. f is the focal length of the two cameras, b is the distance between the two optical centers, also known as baseline and $d = u_R - u_L$ is called as **disparity** is related with projection center of the left camera(p_L)


Object point
Left image
Right image
Projection center of left camera
Projection center of right camera
(a)
P
z
x_l
x_r
c_l
c_r
f
O_l
O_r
T
(b)

Calibration allows for the estimation of f and be values, as well as the calculation of the disparity (d) between two pictures using corresponding points, also known as conjugate points. As seen in Figure (a), Given a point p_L (the projection center of the left image), the dependent image's equivalent point p_L must be found. Due to the tiny difference between the two photos, the matching point can be located in any pixel. Due to the fact that the most frequent similarity procedures investigate each and every point, locating the corresponding point might be a lengthy process. As a result of the epipolar restriction, the search domain can be constrained to a single dimension (along u). According to geometric analysis, the conjugate point of p_L in the second picture must lie on a straight line termed the epipolar line of p_L. Although these two cameras may not be perfectly aligned in more realistic surroundings, it is always feasible to apply a linear transformation to the pictures collected by the camera to accomplish the operation of correspondence selection. This is referred to as the correction.

With help of calibration operation, f and b values can be estimated, the disparity (d) can be calculated with corresponding points also known as conjugate points of two images. As it is illustrated on Figure(a), Given a point (projection center of the left image), the correspondent point p_R in dependent image has to be found. Since the two images are slightly different, the corresponding point can be in any pixel.Finding of correspondent point can be a long progress due to the most common similarity strategies analyze every single point. Therefore, the search domain can be limited to a one dimension (along u) thanks to epipolar constraint.

A geometrical analysis shows that the conjugate point of p_L in the second image, must lie in a straight line called epipolar line of p_L.For more realistic environments these two cameras may not be aligned properly. Nevertheless it is always possible to apply a linear transformation to images which are captured by camera to implement the task of correspondence selection. This operation called as rectification. [11]

4

General Concept of Time-of-Flight Range Camera

In the previous chapter, the Time-of-Flight operating concept has been introduced. Despite its basic simplicity, the real design requires significant effort from all major manufacturers of ToF cameras, including MESA Imaging, PMD Technologies, SoftKinetic, and Microsoft, since measurements require high precision over a few picosecond clock periods. Even a little change in resolution needs multiple clock cycles, which is the time required for the light pulse to traverse that distance back and forth. However, depth measurement accuracy must be managed due to a variety of harmful effects, which may be classified as internal, such as noise or calibration, or environmental.

Diverse methods have resulted in a variety of technologies, while the continuous wave (CW) intensity modulation technique is the most commonly used in commercial solutions. [12] contains information on alternative approaches such as optical shutters (OS) and single-photon avalanche diodes (SPAD).

This chapter presents an overview of ToF cameras and related practical challenges, as well as a discussion of the thesis's primary method, the Time Correlated Single-Pixel Counting (TCSP) system. Additionally, it provides critical strategies for data processing of time-of-flight cameras, including data reduction techniques and solutions to their shortcomings.

4.1 Background

Single-photon LIDAR has established itself as a significant tool for depth imaging in recent years. The technology essentially works by generating a histogram of the time delays between the emission of light pulses and the detection of photon arrivals to determine the depth of a target. Due to the great precision of the time-stamps and the large amount of photons, a significant data processing bottleneck emerges. The complexity of the image reconstruction and the associated factors deteriorate the situation further.

Existing LIDAR approaches have a limiting bottleneck, which may be overcome by constructing a compressive statistic from a sketch of the time delay distribution, which is adequate to infer spatial distance and intensity. [12] Rather than the number of photons or the time-stamp resolution, the size of this drawing scales with the degrees of freedom of the time-of-flight model (number of objects). Additionally, the drawing is well-suited for on-chip online processing.

A histogram may be generated by determining the time delay between the photons sensed by each pixel and the light pulses emitted, as well as the percentage of photons originating from background or ambient light (e.g. the sun).

The number of counts per time histogram bin informs us about the object's depth and reflectivity. The presence of a peak point in the histogram indicates the presence of an object within the range of the LIDAR system. This item is at the same location as the impulsive response. Picture restoration can be used to determine the positions and intensities of the histogram peaks for each pixel in the image if the target item is semi-transparent.

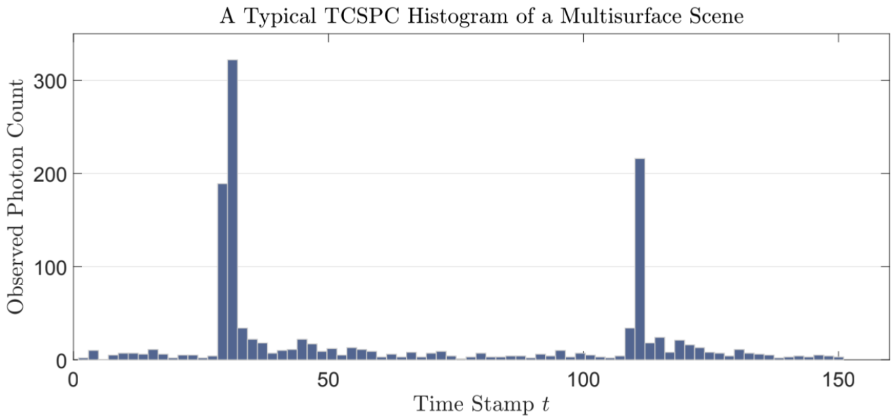


Figure 4.1

For small sketch sizes, it is demonstrated theoretically that the loss of information due to compression is controlled and that the mean squared error of the inference soon converges to the optimum Cramér-Rao bound (i.e. no loss of information). [12] The compressed single LIDAR framework is used to reduce the data in order to build more efficient histograms without compromising any critical information. This reduction may be made up to 1/150 in practice without affecting the reconstructed image's overall resolution.

The process of developing Time-of-Flight image sensors that meet critical requirements such as high rate, high resolution, and low power consumption is difficult due to the massive required data quantities. When the large quantity of photons per pixel and the temporal resolution are combined, this circumstance results in significant data processing issues for the devices. With all of these disadvantages, power consumption as a result of this complexity has a detrimental effect on the data processing problem. Numerous approaches have been proposed to address the trade-off between depth resolution and computational/space complexity. The majority of academic works, such as [13] and [15], present novel perspectives for resolving the conflict between the depth resolution and complexity of the TCSPC histogram. None of these approaches is capable of handling the approximations made on-chip to impair the image's depth resolution. In [14] and [15], a technique is described that gathers the histograms of photon detections during periods of substantial activity. Among all these academic research, only this strategy minimizes data transfer since it is used only at specified points in time.

Nonetheless, these approaches are insufficient when activity is stable and also cannot compensate for the loss of temporal precision caused by narrow histogram binning. Even if [16] introduces a novel way for data reduction during photon transmission, these strategies cannot overcome the disadvantage of repeated histograms of increasing resolution.

The most effective compressive sensing algorithms have been successfully implemented with LIDAR and have concentrated on compressing data across pixels. In [17] provides a strategy for reducing the signal acquisition costs by using the sparsity of natural sceneries in a certain representation domain (e.g wavelet transform). The depth accuracy is limited by the size of the amplitude noise and the decay rate of the impulse response. Therefore, this constraint is associated with a single surface per pixel. In a similar vein, [18] suggests a scene-dependent adaptable sampling technique. By iteratively recreating regions of interest and depth maps derived from data, the reduction ratio of predicted circumstances may be increased to eight-fold.

However, these approaches may be used to compress data inside the spatial domain, rather than across the depth or time domain, as the method used for the single pixel detection component does, and are thus fundamentally different in practice.

4.2 Time Correlated Single Photon Counting

Time-correlated single-photon counting (TCSPC) is a common technique to measure impulse decays in the time domain. In the core case, single-photon events are detected and their arrival time is correlated to the laser pulse which was used for the excitation of the sample.

In this section, theoretical structure of the TCSPC has been introduced with associated expressions.

4.2.1 Data Distribution

A Poisson distribution is used to represent the photon count at time-stamp $t \in [0, T-1]$ for every given pixel.

$$y_{tk}|(r, b, t_k) \sim P(rh(t - t_k) + b) \tag{4.1}$$

where $r \geq 0$, r signifies the measured surface's reflectance, $h(.)$ the system's impulse response, and b denotes the intensity of background photons. T is the number of discretized time-stamp bins throughout the range of interest. Depending on the time-stamp resolution Δt, the time-stamp t is given in the range $[0, T-1]$. To clarify, despite the fact that the distribution may handle more complicated scenarios, we can assume that the integral of the impulse response $H = \sum_{t=0}^{T} h(t = 1)\Delta t$ is stable. [12]

Furthermore, in [19], a mixture distribution is provided, from which the arrival time of the observed pth photon may be modeled. Assuming K different reflecting surfaces, we may state that α_k and α_0 denote the probability that the detected photon originated from the kth surface and background sources, respectively. If we indicate the time-stamp of the pth photon as $x_p \in [0, T-1]$ where $n \geq p \geq 1$ is a mixture distribution, then x_p may be described by a mixture distribution.

$$\pi(x_p|\alpha_0, ..., \alpha_K, t_0, ..., t_K) = \sum_{k=1}^{K} \alpha_k \pi_s(x_p|t_k) + \alpha_0 \pi_b(x_p) \tag{4.2}$$

The uniform distribution $\pi_b(x_p) = 1/T$ and the $\pi_s(x_p|t) = h(x_p - t)/H$ define the distribution of photons arisen from signal and background. In practice, the signal distribution π_s is generally represented by either a discretized Gaussian distribution spanning the range $[0, T-1]$ or a data-driven impulse function derived through research and experiments.

4.2.2 Statistical Evaluation

In the previous subsection, we discussed mixture distribution in the context of time-stamp photon detection. The parameter estimation is based on the set of parameters $\theta \in \Theta \subset R^{2K+1}$ associated to the probability model $\pi(., \theta)$ which is specified in a space $x \in R^d$. The dimension d refers to the situation of a single-photon LIDAR. In Maximum likelihood estimation (MLE), the finite data-set $X = x_{i_{i=1}}^{n}$ of n samples, which we assume is sampled and observed on a regular basis, is one of the most common parameter estimation methods from the given distribution, and a likelihood function associated with the finite data is maximized with respect to the model parameters.

$$\theta = arg_{\theta} min \frac{1}{n} \sum_{i=1}^{n} log \pi(x_i|\theta) \tag{4.3}$$

1) Generalised Method of Moments: In some cases, the likelihood function might not have a closed form solution nor a computationally trackable approximation. The generalised method of moments [20], [21] (GeMM) is a technique for estimating parameters that involves matching a collection of generalised moments to their empirical counterparts computed over a set of finite data drawn from the distribution $\pi(x|\theta)$. Given a nonlinear function $g : R^d \rightarrow C^m$, then we define the expectation constraint

$$\mathbb{E} g(x; \theta) = 0 \tag{4.4}$$

where E is the expectation of the probability distribution $\pi(x|\theta)$. To try to enforce the moment constraints of (4), the GeMM estimator is often derived by minimising a quadratic cost of the empirical discrepancy with respect to θ. Let us begin by defining

$$g_n(X; \theta) := \frac{1}{n} \sum_{i=1}^{n} g_i(x; \theta), \tag{4.5}$$

GeMM can be stated in the form shown below if it is determined for $X = \{x_i\}_{i=1}^{n}$.

$$\theta = argmin_{\theta} g_n(X; \theta)^T \mathbf{W} g_n(X; \theta) \tag{4.6}$$

where **W** is a symmetric positive definite weighting matrix that depends on θ.

2) Compressive Learning: Compressive learning [22], [23] builds on the notion of GeMM by utilizing generalised moments of data, but with the unique purpose of lowering signal acquisition, spatial, and temporal complexity. The connection to GeMM is formed by dividing the function g into the following form:

$$g(x;\theta) = \Phi(x) - \mathbb{E}_\theta \Phi(x) \tag{4.7}$$

where $\Phi : \mathbb{R}^d \longmapsto \mathbb{C}^m$ is commonly referred to as the feature function. The separable form dissociates the measured moment, $\Phi(x)$, from the parameters θ to be estimated. This is not a common assumption in GeMM, but it may occur in rare circumstances. By referring to the empirical mean or so-called sketch as

$$z_n := \frac{1}{n} \sum_{i=1}^{n} \Phi(x_i) \tag{4.8}$$

It can be approximated with help of θ using the sketch z_n by minimising

$$\theta = argmin_\theta ||z_n - \mathbb{E}_\theta \Phi(x)||_w^2 \tag{4.9}$$

which is the specific compressive GeMM loss of (6). In Section III, we define the weighting matrix W for compressive single-photon LIDAR directly.

The separable form of g in (7) enables the formation of a sketch statistic z_n with a single pass of the data without the need to keep X, and it can be quickly updated on the fly with minimum computational expense. The sketch statistic has size m, or size *2m* if separated into its real and imaginary components, and scales essentially irrespective of the dataset X dimensions, which in the case of single-photon LIDAR are the photon count n or the binning resolution T.

4.2.3 Compressing Single Depth Data

The sample mean of all photon time-stamps ($\Phi(x) = x$) is the simplest summary statistic for estimating the single location parameter t_1 in the absence of photons from background sources and the presence of a single surface or object. This is only true in the noiseless scenario, since when background photons are detected, the sample mean estimation is substantially inclined toward the center of the histogram.

Assume instead that we measure the cosine and sine of each photon count x x with angular frequency $w = \frac{2\pi}{T}$,

$$\Phi(x) = \begin{bmatrix} cos(\frac{2\pi x}{T}) \\ sin(\frac{2\pi x}{T}) \end{bmatrix} \tag{4.10}$$

where z_n is the real valued sketch of size $2(m=1)$ generated across the dataset X as seen in (8). The trigonometric sample mean may be used to obtain an estimate of the single depth location parameter t_1 straight from the drawing, without using the data X.

$$t_1 = \frac{T}{2\pi} arg\left\{ \sum_{j=1}^{n} cos\left(\frac{2\pi x_j}{T}\right) + i \sum_{j=1}^{n} sin\left(\frac{2\pi x_j}{T}\right) \right\} \tag{4.11}$$

where *arg* denotes a complicated argument. Because the background photons are spread equally over the $[0, T-1]$ range.

If the preceding is summarized using a simulated example, where a pixel of T = 1000 histogram bins with a signal-to-background ratio (SBR) of 1 and a total of n = 600 photons is simulated, with the time-stamp of each photon given by $X = \{x_i\}_{i=1}^{n}$. The data was simulated using a Gaussian impulse response function with $\sigma = 15$ and a real position of time-stamp with $t_1 = 320$; the drawing is shown below. The graphic depicts both standard and circular mean estimations. In particular, circular mean estimation can suffer from the noise effect in TCSPC illustration.

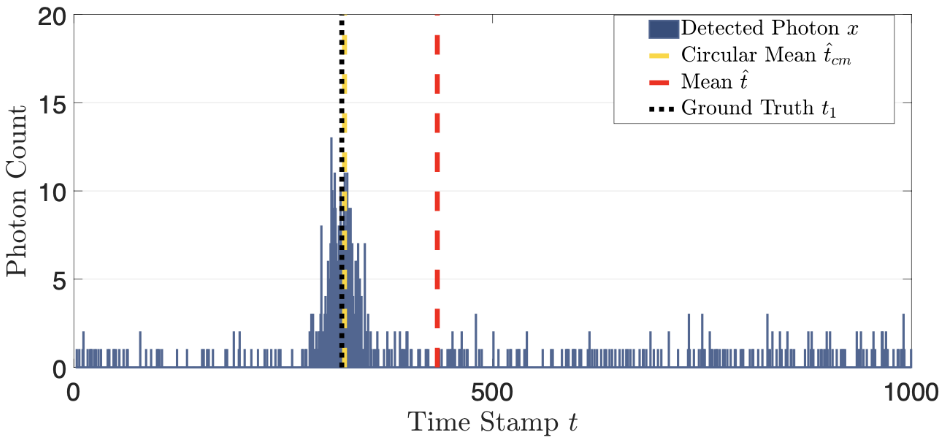


Figure 4.2: Illustration of Photon Detection and the Ground Truth

The TCSPC histogram is shown with $t_1 = 320$. The circular mean estimate (yellow) is placed on the standard mean estimate (red).

The Single Photon Counting section is based on a simulation of a synthetic dataset generated from this technical information. in figure 4.2

Next the technical major phases of Time Correlated Single Photon Counting, the following sections will explore distance prediction using classical methodologies and machine learning techniques.

4.3 Pile-up Effect

The Pile-Up effect outlines the consequences of photons lost due to the TCSPC devices' dead period at high photon count rates. There is a dead period in most single photon counting detectors and TCSPC circuits. Following the detection of one photon, the device requires some time to prepare for the detection of the next photon.

Let's use a 40 MHz laser repetition rate as an example. A laser pulse is delivered to the fluorescence molecules every 25 nanoseconds. After one laser blast, we can only detect one photon due to the dead period. The term "pile up" refers to the impact of photons wasted based on the TCSPC devices' dead period at high photon count rates. There are two effects:

- The average measured lifespan decreases.
- With the inclusion of a shorter component, a mono-exponential decay becomes bi-exponential.

The pile-up effect must be avoided since it distorts the histogram. To do so, the count rate monitored should not surpass a particular threshold in regard to the laser repetition rate.

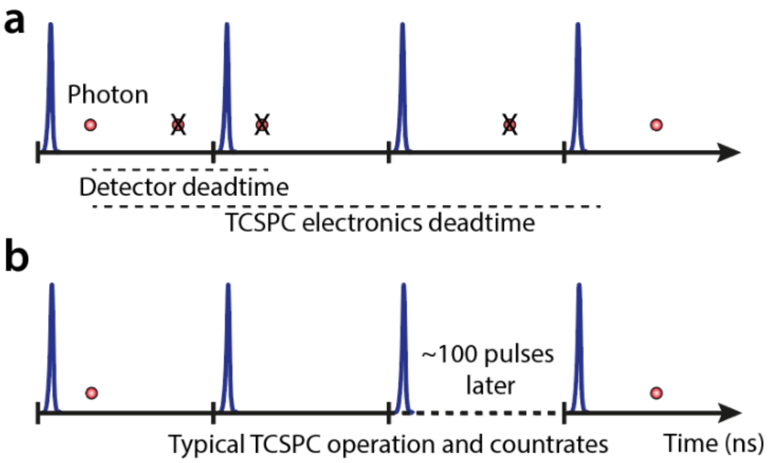

a
Photon
Detector deadtime
TCSPC electronics deadtime
b
~100 pulses
later
Typical TCSPC operation and countrates
Time (ns)

Figure 4.3: Illustration of Pile-up Effect

(a) When photon collection rates are high in TCSPC, the instrument dead periods ends with the loss of information (X) of photons that are arriving late. The probability density function is inclined towards shorter lifespan values, resulting in a decrease in photon efficiency.

(b) To eliminate photon pileup effects, most TCSPC systems use low laser intensities and low photon count rates (1 photon/100 excitation pulses). As a result, image detection times are in the order of minutes.

4.4 TCSPC-Based Distance Estimation

Distance Estimation with histogram dataset have many difficulties due to dimension (iteration number:10, bin:20 , histogram points:1024) histogram size for every distance sample) of the dataset. To provide a sufficient model can be possible help of higher epoch counts, on the other hand, there is a trade-off between performance and high accuracy. The approach makes a comparison among more traditional distance estimation process such as threshold based estimation and most common ML approaches which are exploited for prediction task.

To reduce the data dimension for histogram bins, the first approach, threshold calculation. Due to real laser light is reflected from target in t time, the biggest magnitude of the histogram bin should remarks the time (t). With help of this logic first the method calculates maximum and minimum values of the histogram, in which time slots have the highest and lowest photon count, and then to reduce the 1024 points, the average value is taken as a threshold value and then we can eliminate the half of histograms, without accounting them to the estimation. Before the training step, this preprocessing diminishes the dataset effectively

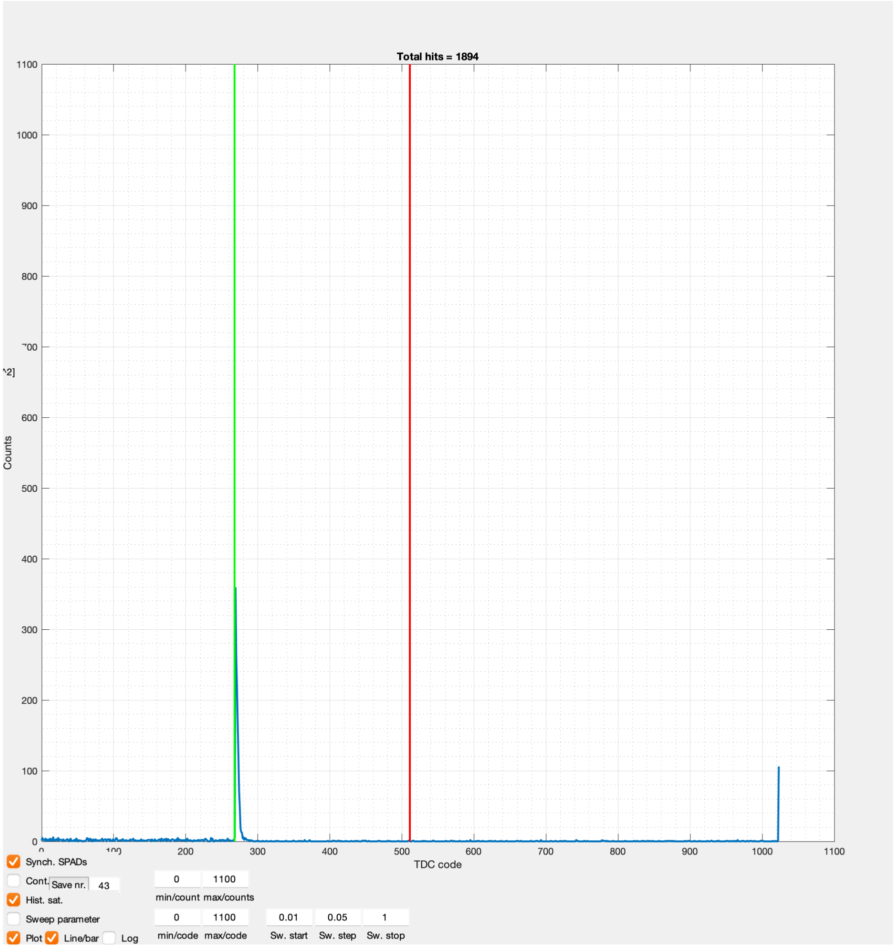
In addition, this approach can be exploited directly for distance estimation with inaccurate results. The peak point of the threshold value is considered as index value of the histogram and then with speed of light, it gives the distance estimation. Moreover, half time should be accounted for this calculation and the system works in picoseconds. Especially, shortest distances can be calculated with help of this approach. Nevertheless, to handle more robust results, ML approaches should be implemented.

For the main approach, the base distance which is created with data of histogram is used, This histogram has the lowest distance which can be detected by the time-of-flight sensor (e.g=0). With optimal target reflectivity and base distance are exploited as a reference point for the similarity calculation. Deep Regression ML algorithm is implemented for this task as a single and multi layer architectures, instead of more complicated models, this method provides us more acceptable running time and results for this dataset. To obtain a model which can be exploited in the real time or preferably real time is the most important criteria for the project.

Therefore, single layer deep regression with *feature number: 1024* , *hidden neuron: 400, output neuron: 1* is implemented. Furthermore, multi layer deep regression with same hyper-parameters is implemented to observe the results and make a comparison for the best combination. The optimizer is selected SGD with learning rate: 10^4.The methods are analyzed for the analysis which will be explained in Results chapter consider the base distance.

4.5 Dataset

For Time Correlated Single Photon Counting task, the simulation which is provided by Fondazione Bruno Kessler, IRIS Research Team was exploited.



☑ Synch. SPADs

□ Cont. Save nr. 43

☑ Hist. sat.

□ Sweep parameter

☑ Plot ☑ Line/bar □ Log

min/count 0 max/counts 1100

min/code 0 max/code 1100

Sw. start 0.01 Sw. step 0.05 Sw. stop 1

Figure 4.4: A histogram sample which is taken from simulation

The figure 4.4 shows a histogram of ToF simulation which shows photon counts during period. Horizontal line shows histogram time stamps [0,1024], the vertical line shows number of total hits which are recognized by sensor. The green light illustrates the estimated histogram bin for the distance is selected and red line shows average point of the horizontal line. The dataset comes to vision with distance histograms which are collected from this simulation tool. This tool provides us synthetic time of flight histogram values, with changing hyper-parameters to obtain a model which satisfies all conditions. For this reason, the dataset created with images between 1-10 meters (the max flight range: 15.27 meters) with different reflectivity and optical transmittance parameters. For this dataset, as it was explained in the previous sections, dead time is considered as $100ns$, and then other parameters ; number of histogram points: 2000, laser divergence: 0.18° , laser peak power: $0.3W$. The dataset contains histogram inputs from 1 meter to 10 meters for every 0.1-meter distance. As a part of the supervised approach, all data is labeled by its distance.

5
Stereo Vision System

In spite of stereo devices are produced by several companies , additionally, two single standard cameras can be exploited for this task. On the other hand commercial stereo products may be considered as a base point for solid framework. Baseline, focal length, calibration and synchronization issues are handled by the devices properly. Source code , Software Development Kit (SDK) including drivers, additional libraries, programs ,and interfaces are provided for integration process. Nevertheless, using two single cameras provides more flexibility for the users, in particular selection of parameters on the system is easier.

In this chapter, before describing the procedure used to compute disparity map and depth maps for stereo devices, more details on stereo algorithms are discussed. Especially, essential issues of correspondence selection, or disparity computation, are analyzed in order to understand the reasoning behind depth estimation.

5.1 Stereo Matching Algorithms

The purpose of stereo matching algorithms is to connect points in one picture to corresponding pixels in another image based on specific criteria.

- *Similarity:* it is implied in the correspondence issue that the points on both pictures must be comparable.

- *Epipolar geometry:* By definition, the projection of P onto the second image must be on the second image's epipolar line. Thus, using a common idea of epipolar geometry, we may establish a strong constraint between picture pairings without understanding the scene's three-dimensional structure. As previously stated, the conjugate point is located in a straight line.
- *Smoothness:* Smooth surfaces have a constant depth away from the edges.
- *Uniqueness:* Each point in one image must match to a single point in the next image. This assumption can be disproved if transparent things exist.
- *Monotonic order constraint:* If a point p_1 in one picture corresponds to p_1' in another, the counterpart of another point p_2 to the right (left) of p_1 must be p_1'. This criterion is violated if p_2 is located between the optical centers of the two cameras in a specific conoid region represented by p_1.

The majority of academic works on stereo algorithms, for example Scharstein and Szelisky [27], examine the following basic implementation blocks: computation, cost aggregation, matching cost, disparity computation, and disparity refining.

The first step is to determine the cost of matching. The most often used techniques for pixel-based matching costs are the sum of squared differences (SSD), sum of absolute differences (SAD), normalized cross correlation (NCC), and census transform. This strategy can benefit from a variety of pre-processing techniques, including Laplacian, Gaussian, and bilateral filtering. Calculating the mean value of the window may be useful for reducing noise and photometry distortion. In both local and window-based systems, cost aggregation is accomplished by summing or averaging throughout a support zone. The support zone may be two-dimensional in the simplest cases or three-dimensional in more complex cases to provide additional support for slanted surfaces. Aggregation may be achieved well by convolution or box-filtering.

Local and global approaches can be used to compute disparity. In the academic study [27], another strategy is included that is based on semi-global approaches, which is similar to dynamic programming and cooperative algorithms. Local approaches may be thought of as those that identify only similarities between the region surrounding a pixel and similar-shaped regions around all possible conjugate points on the other picture. The window's dimension can be either stable or unstable in order to get more robust findings for all places in the picture. The disparity is chosen in accordance with the approach of maximizing of similarity (Winner Takes All).

As will be demonstrated in the next sections, local approaches cannot meet all major stereo vision expectations. As a result of the lack of regularization, the findings are not free of noise. [25] discusses a possible solution to this problem using a weighted box filter and the cost volume. Another significant issue is the "edge-fattening effect," which occurs as a result of aggregation across a support window in stereo systems. This issue can be resolved using the approach outlined in [24]. This technique protects the edges more well than bilateral filters, and its execution time is not dependent on the filter size.

On the other side, global approaches do not focus on individual point pairings, but they may estimate all disparity values concurrently when global optimization schemes are used. The basic principle is to design a disparity function (d) that minimizes the global energy generated by a component that quantifies how well the disparity function agrees with the input picture pair. Bayesian formulations are a frequently used strategy in global approaches. These techniques, in particular, treat the entire scene as a Markov random field (MRF) and entail the calculation of unique framework values based on local comparisons between the two pictures and scene depth smoothness requirements.

Additionally, after global approaches, semi-global methods are used to estimate discrepancy. To decrease complexity, the cost function is minimized using a reduced model for all points in the disparity picture, as opposed to global techniques that estimate the whole disparity image. In contrast to global strategies, it does not attempt to evaluate the whole gap simultaneously. For example, the two most well-known semi-global approaches, Dynamic Programming and Scanline Optimization, operate in the one-dimensional domain and optimize each horizontal picture row independently. Semi Global Matching (SGM) is a more advanced version of the semi-global stereo method.

Regardless of the fact that new methods are being developed to achieve a more robust solution to the correlation problem, the quality of stereo reconstruction is fundamentally dependent on the scene features.

Besides, another widely used approach method can be used to improve stereo matching computations, particularly in uniform areas. The system may be utilized with the assistance of an external lighting device. Active stereo requires two cameras, and this external light aids with correspondence selection.

5.2 Keypoints for Matching

The detection of pairs of conjugate pixels is considered the most difficult phase in the calculation of the depth map. This is, in general, one of the most difficult problems in computer vision. The essential premise is that the correspondence problem is based on slightly different images that must exhibit a certain amount of discrepancy. Fundamentally, the majority of issues are driven by correspondence detection, which cannot be improved while the baseline increases. Nonetheless, a sizable baseline is necessary to achieve a statistically meaningful disparity. The following sections discuss the key difficulties associated with correspondence selection:

- **Occlusions and discontinuities:** Due to the discontinuities in the surfaces and the object's specific displacement in the scene, some of the points in the first picture may not correspond to the points in the second image. For such locations that lack the relative conjugate, there is no rationale or meaning to characterize the disparity. This is the most prevalent problem in stereo vision and may be detected by first gazing at the edge of an item and then recognizing it with one of the two eyes. To recognize detection occlusions, a process called Left-Right consistency check might be used. Nonetheless, there is no adequate answer for resolving the gap in these areas.

- **Radiometric distortion and noise:** Additionally, for Lambertian materials, the observed spots may not be same in these two pictures. Due to the presence of noise, the hue and intensity of two scenes may differ, complicating the correspondence search.

- **Specular surfaces:** As with the last item, glossy materials can reflect light directly into the camera. Due to the two cameras' differing view angles, a section of the picture may be visible while the same location in the other image may be overexposed. When the scene's illumination system does not rely on a direct spotlight, the probability of overexposed sections reduces.

- **Perspective foreshortening:** Due to the somewhat diverse views provided by each stereo camera, the picture of the surface can be compressed and occupy a smaller area in one view. The effect is more obvious when an item is horizontally tilted. Foreshortening has disadvantages, particularly for approaches that aggregate costs using fixed-size windows, because it is expected that objects fill the same extents in both pictures

- **Transparent objects:** Transparent objects have an inherent ambiguity. These things would obscure or conceal the true background. Without a doubt, this circumstance has a detrimental effect on both local and global methods, lowering their performance.

- **Uniform regions:** The majority of stereo matching methods are incapable of dealing with poorly textured surfaces in the picture. Neither global nor local techniques are enough for resolving this issue. The detection of comparable regions is dependent on the function's peak point, which is obtained using correlation or another approach. Despite the fact that this is a typical limitation of all stereo matching methods, strategies are capable of assigning a suitable disparity to all of these places.

- **Repetitive pattern:** Regions devoid of texture or densely textured with periodic patterns provide us with several obstacles when creating a scene. Due to a lack of global information about the scene, it is extremely difficult to tell the difference between the true correspondence and fake versions. The most often used technique is a checkboard with a cost function for the points. Additionally, in this scenario, ambiguity can be eliminated by the use of global techniques.

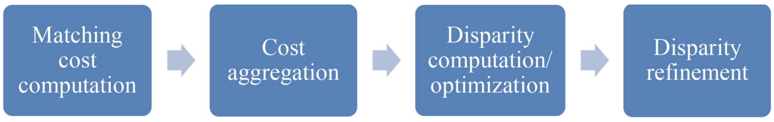

Matching cost computation
Cost aggregation
Disparity computation/ optimization
Disparity refinement

Figure 5.1: The General Flow Chart of Disparity Map

5.3 Depth estimation from stereo vision

Numerous suggestions for disparity computation have been published in the literature. Among these academic papers, Hirschmuller's Semi-Global Matching (SGM) was considered the best option. It accurately depicts the scene's three-dimensional structure using a point-wise matching cost and a smoothness term. The majority of one-dimensional energy functions computed using distinct pathways are individually and effectively reduced, and their costs are added together. Eight to sixteen distinct separate pathways are utilized in academic investigations that suggest SGM. The discrepancy corresponding to the lowest aggregated cost is chosen for each point.

OpenCV [26], one of the most widely used computer vision libraries, implements a modified version of this approach in an optimal manner. It differentiates itself by utilizing a different matching cost computation. Instead of the original mutual information cost function, the Birchfield-Tomasi sub-pixel metric is utilized.

Disparity maps can be exploited with data fusion frameworks for estimation. In previous studies which involve stereo matching algorithms remark that Semi Global Matching is one the of the most efficient and fastest approach for this task.

5.4 Depth Estimation with Neural Networks

After the explanation of more traditional approach for disparity maps and key-points of this approach, this part introduces deep learning strategy for disparity maps as a catalyse for performance and results.

As it is illustrated in the figure, the flow chart is based on three main steps,

- (1) view-feature extraction
- (2) feature fusion
- (3) disparity regression

All these steps are merged with each other without any additional process. In the next sections, these steps will be discussed.

5.4.1 View-feature extraction

Two sub-networks of a 2-D CNN are utilized to extract features from each view picture, as seen in Figure 5.2 To begin, 5x5 convolutional filters with stride 2 are used to capture more global information and build a 3-D feature map with half the resolution. The last layer is a 3x3 convolutional filter, followed by eight residual blocks composed of two 3x3 convolution layers each. Each convolutional layer is followed by a ReLU activation layer to induce non-linearity. The two sub-networks share the weights for feature extraction from the left- and right-view pictures.

5.4.2 Building 4-D disparity-varying feature volume

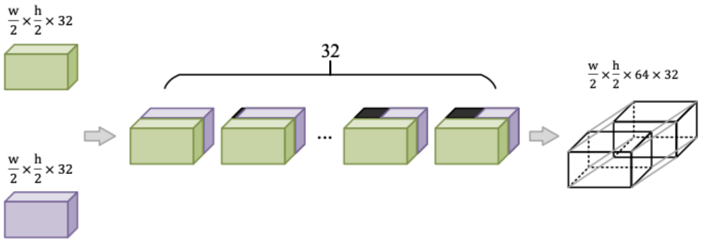

$\frac{w}{2} \times \frac{h}{2} \times 32$
32
$\frac{w}{2} \times \frac{h}{2} \times 64 \times 32$
$\frac{w}{2} \times \frac{h}{2} \times 32$
...

Figure 5.2

Instead of using the original RGB intensities, the disparity information can be generated by integrating the two 3-D feature maps taken from the left- and right-view images. It is tried to concatenate the two feature volumes from Fig.5.2 across different disparity levels and then pack them into a 4D volume, as suggested by [30], rather than merging them by typical direct concatenation. The network will now be able to learn semantics as a result of this.

If it is assumed that 32 different disparities and 31 horizontal shifts (i.e., in the w axis) for the right-view features with zero padding, as illustrated in figure which is given the above. (the black area in Fig. 5.2). (w/2)x(h/2)x64x32 is the size of the final 4-D volume.

5.4.3 Disparity regression

To regress the discrepancies from the 4-D feature volume, a 3-D CNN based on a modified UNet [29] network (improvement from UNet) is utilized, as shown in Fig. 5.3. The original U-Net employs a pyramid encoder-decoder design, with symmetrical encoder (i.e., contraction path) and decoder (i.e., expansion path). In a pyramid, there are numerous levels (scales), with reduced resolution at higher levels (i.e., near the pyramid peak in Fig 5.3) but a greater number of channels. To achieve high accuracy, the encoder uses down-sampling to extract image features at various scales, while the decoder predicts the class or output from the extracted features by up-sampling and fusing the features at higher scales.

[29] improves UNet by re-scaling the direct connection between the encoder and decoder paths. To fuse features at hierarchical scales, image features at a scale will be concatenated with up-sampled features at higher scales before being fed into the decoder path. Before feeding each CNN layer, UNet concatenates features with those generated from other CNN layers at the same scale. The modified UNet network includes four tiers in the pyramid, with 32, 64, 128, and 256 channels in the output feature map, respectively. After feeding the 4D feature volume into the regression sub-network, a succession of 3D convolutional layers, down-sampling,concatenation, up-sampling, and skip connection are used to create a 4-D feature volume of size $wxhx6x32$ that is then output. In addition, each convolutional layer employs a $3x3x3$ convolutional filter with padding of 1. Down-sampling is done via max-pooling, and upsampling is done with bilinear interpolation.

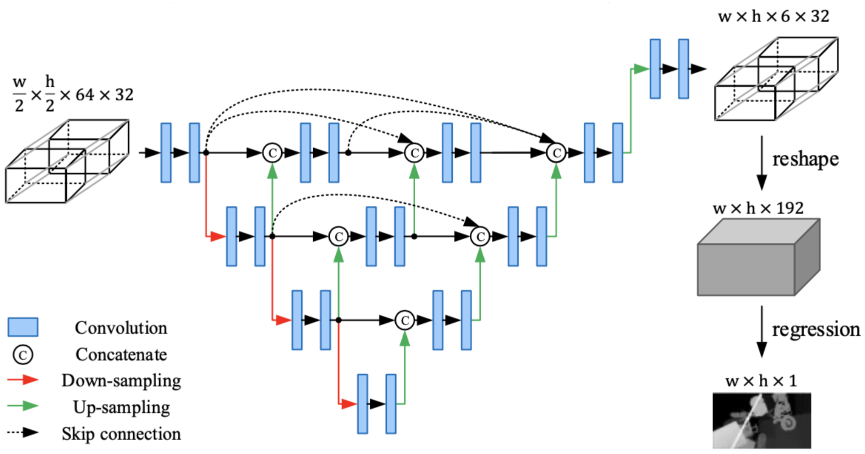


Figure 5.3

By concatenating the 32 3-D feature volumes in channel dimension, the output of UNet++'s 4-D feature volume (wxhx6x32) is transformed to wxhx192. For each pixel in the left-view image, this refers a partition into 192 possible disparity values. The disparity for each pixel is then estimated using a soft-argmax function, which produces the final disparity map. The output disparity $\hat{d}$ can be computed as:

$$\hat{d} = \sum_{d=0}^{D_{max}} x\sigma(c_d) \tag{5.1}$$

where c_d is the possible disparity category, d is the corresponding disparity value, σ is the softmax function and D_{max} is set to 191. [28]

5.4.4 Loss function

Before training, the model parameters are randomly initialized. After that, an end-to-end supervised learning is carried out. Because some disparity ground truths are formed from 3D point clouds, the disparity map that results may contain sparse pixels when contrasted to the RGB input image. Only pixels with disparity ground facts are taken into account in the loss function, which is defined as:

$$L(d, \hat{d}) = \frac{1}{N} \sum_{n=1}^{N} ||d_n - \hat{d}_n||_1 \tag{5.2}$$

where d is the ground truth disparity $\hat{d}$ is the estimated disparity, and N is the number of labeled pixels.

6

Monocular Depth Estimation

Estimating the depth of a scene from a single image is an easy task for humans, but is notoriously difficult for computational models to do with high accuracy and low resource requirements. Monocular Depth Estimation (abbr. as MDE hereafter) is this task of estimating depth from a single RGB image. Some applications include scene understanding, 3D modeling, robotics, autonomous driving, etc. Recovering depth information in these applications is more important when no other information such as stereo images, optical flow, or point clouds are unavailable. Monocular depth estimation is often described as an ill-posed and inherently ambiguous problem. Estimating depth from 2D images is a crucial step in scene reconstruction. The problem can be framed as: given a single RGB image as input, predict a dense depth map for each pixel. This problem is worsened by the fact that most scenes have large texture and structural variations, object occlusions, and rich geometric detailing. All these factors lead to difficulty in accurate depth estimation.

In this part, Monocular Depth Estimation will be introduced with main details and theoretical background after that, the neural network approach for depth estimation which is based on supervised modelling will be discussed.

6.1 Definition of Problem

The basic formulation of monocular depth estimation require this concept: RGB-depth (RGB-D) pair images which are obtained from single camera resource. And then, monocular depth estimation from single images may be considered as a pixel-level continuous regression problem. This regression problem, is a standard mean square error (MSE) loss in log-space or its variants are used as loss function. This approach is limited by supervised learning in which the task is pixel-wise continuous regression.

If more details are given with a theoretical expression, it can be illustrated in below, Let I be the space of RGB images and D the domain of real-valued depth maps. Given a training set $T = \{(I_i, D_i)\}_{i=1}^{M}, I_i \in I$ and $D_i \in D$, the task is to learn a non-linear mapping $\Phi : \mathbb{I} \longrightarrow \mathbb{D}$. This formulation is applicable to supervised learning algorithms where pixel- level ground truth is available. Some methods relax this constraint by introducing different requirements and constraints. [53]

6.2 Depth Map Prediction using a Multi-Scale Deep Network

Previous academic studies which are based on depth estimation, focus on Convolutional Neural Networks with multi-scale features. The introduction of this approach to use multi-scale information was introduced in this part . This network has two main components as one that first estimates the global structure of the scene, after that a second which improves this estimation with using local information. They use a specific scale-invariant loss to calculate the scale dependent error.

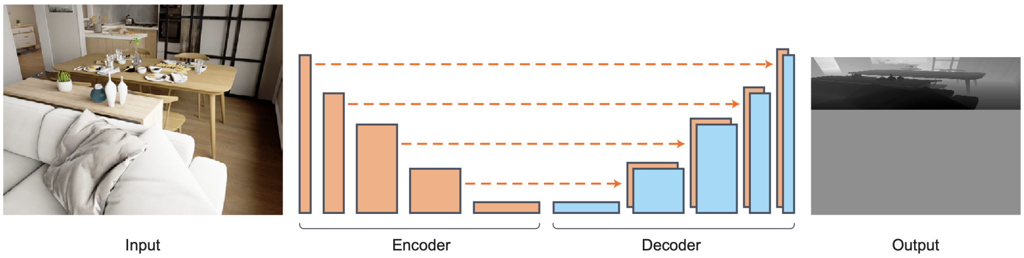


Figure 6.1: General Input - Output Sequence

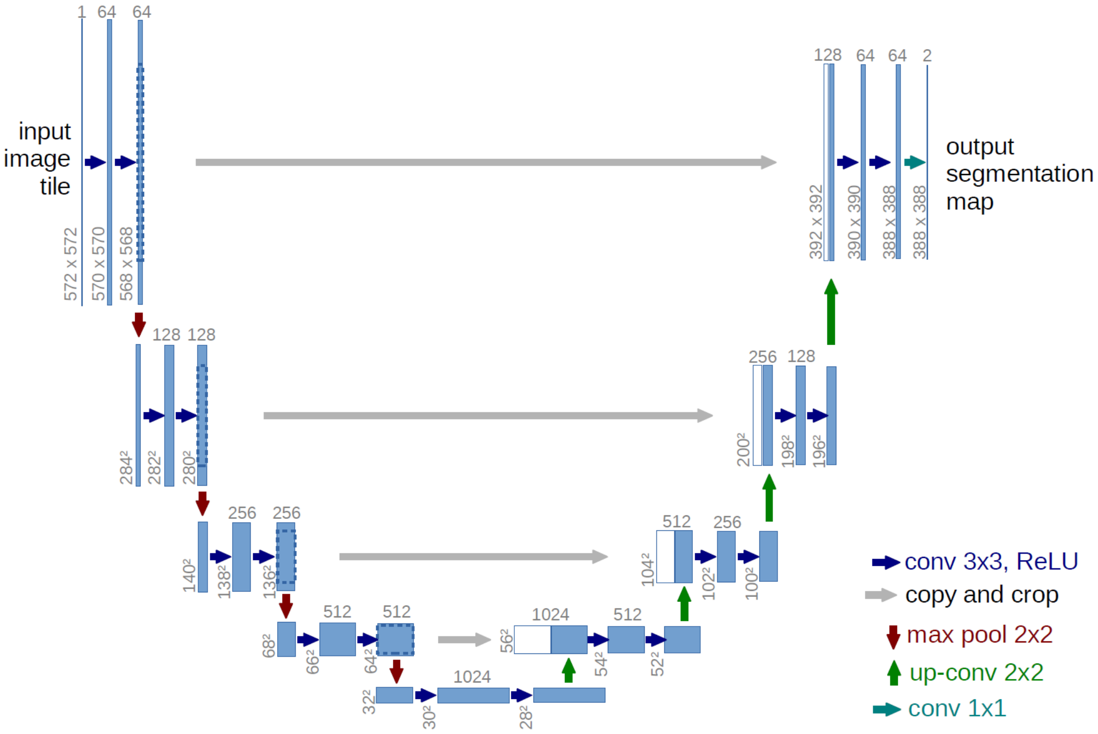


Figure 6.2: U-net Architecture

6.3 U-net

U-net is a type of convolutional network architecture that is used to segment pictures quickly and precisely. It has outperformed the previous best approach (a sliding-window convolutional network) up to this point.

In spite of Convolutional Neural Networks expose effective results for image segmentation and depth estimation issues, for complex datasets, they cannot be as good as U-net architecture. First, it was designed for medical image segmentation and then different usage areas came to vision.

The schema which is given the above illustrates the general concept of this approach with sample layer dimensions.

6.4 Network Architecture

- **Architecture:** The figure demonstrates the depth estimation encoder-decoder network in its entirety. The U-net architecture for this part makes use of pre-trained ImageNet [53] weights and a feature vector encoded by the DenseNet-169 [44] network. Then, in order to reconstruct the final depth map with a resolution half that of the input, this vector must be passed through a series of upsampling layers [48]. The decoder creates these upsampling layers and associated skip-connections. The decoder also supports Batch Normalization, however the design does not contain the extra advanced layers recommended in current state-of-the-art techniques [41], [43].

- **Complexity and performance:** The performance of this surprisingly simple design prompts further inquiry into the components that contribute most to the production of these high-quality depth maps. Diverse state-of-the-art encoders [38] with greater or lesser complexity than DenseNet-169 were seen, as well as a variety of decoder approaches. Not only do more advanced encoder-decoder models give less trustworthy outputs, but they also operate more slowly. As a result, they have no effect on performance. Simple bilinear structures are utilized to combine upsampling steps effectively, resulting in increased performance.

- **Encoder-Decoder Networks:** Encoder-decoder structure is integrated to many deep network models for depth estimation [32], [33], [34], [35], [36]. According to one formulation, the U-Net architecture may create skip connections between convolution layers on the encoder path and upsampling layers on the decoder path that have the same spatial dimension. All of these connections between feature maps are employed to recover and enforce spatial information across several resolutions and to maintain spatial consistency on the output picture, which requires alignment of the input and output channels. [31] According to previous research [37], introducing an adversarial term boosts high-level information while maintaining object borders and shape details. The implementation of the U-net is based on 256 x 256 input images that are sampled down to 1x1 pixel. Up-sampling and pooling are eliminated in favor of 4x4 convolution filters with stride 2x2 and transposed convolutions.

6.5 Learning

Loss Function: The usual loss function for depth regression concerns the difference between the true depth map (y) and the depth regression estimation ($\hat{y}$). Different loss function techniques can significantly impact the training time and overall performance of depth estimation. The major part of loss function adjustments employed in neural network optimization are documented in the literature on depth estimation [40], [47], [50], [41]. This approach proposes a loss function that finds a balance between recreating depth images and reducing the gap between depth values while penalizing distortions of high frequency features in the depth map's image domain.

These details are typically corresponded to the boundaries of objects in the scene. For training the network, the loss L between y and ŷ can be expressed as the weighted sum of three loss functions.

$$L(y,\hat{y}) = \lambda L_{depth}(y,\hat{y}) + L_{grad}(y,\hat{y}) + L_{SSIM}(y,\hat{y}) \tag{6.1}$$

According to academic study [53], the first loss term L_{depth} is the point-wise L1 loss defined on the depth values:

$$L_{depth}(y,\hat{y}) = \frac{1}{n}\sum_{n}^{p}|y_p - \hat{y}_p| \tag{6.2}$$

In addition, same academic study remarks a second loss term L_{grad} is the L1 loss defined over the image gradient g of the depth image:

$$L_{grad}(y,\hat{y}) = \frac{1}{n}\sum_{n}^{p}|g_x(y_p,\hat{y}_p)| + |g_y(y_p,\hat{y}_p)| \tag{6.3}$$

where g_x and g_y, respectively, compute the differences in the x and y components for the depth image gradients of y and $\hat{y}$. Lastly, L_{SSIM} uses the Structural Similarity (SSIM) [51] term which is a commonly-used metric for image reconstruction tasks. It has been recently shown to be a good loss term for depth estimating CNNs [42]. Since SSIM has an upper bound of one, we define it as a loss L_{SSIM} as follows:

$$L_{SSIM}(y,\hat{y}) = \frac{1 - SSIM(y,\hat{y})}{2} \tag{6.4}$$

As another additional information, this academic study suggests to define one weight parameter λ for the loss term L_{depth}. It was empirically found and set $\lambda = 0.1$ as a reasonable weight for this term.

One of the most common drawbacks in loss functions is when the ground-truth depth values are bigger, also loss terms may tend to be larger. To solve this issue, disparity [50], [45] where for the original depth map y_{orig} is considered for calculation. If y refers to the target depth map, $y = m/y_{orig}$ where m is the maximum depth in the scene (e.g. m = 10meters for the NYU Depth v2 dataset). Nevertheless in this case, the loss value gets larger for smaller values.

Augmentation Policy: Some pre-processing steps can improve the results rapidly especially to handle the different data structures in the images lead to over-fitting. In [46] considers geometric and photometric transformations for generalization performance. Due to the network is designed to robust results for every image in the dataset, only geometric transformation cannot be sufficient in all cases. Vertical and horizontal flips, and image rotations can be considered for the network. Among these strategies, exploiting the image rotation and right left flip are the bests choice for the NYU Dataset. Therefore, these operations are applied. Photometric transformations handle different color channels on the input and can increase the performance. Furthermore, to increase the efficiency, this feature was integrated to the network.

6.6 Implementation

6.6.1 Implementation Details:

The implementation of proposed depth estimation network using TensorFlow and trained with 16GB memory. The encoder is selected as a DenseNet- 169 pretrained on ImageNet. The weights of the model for the decoder part are randomly initialized. In this model, the ADAM optimizer is exploited with *learning rate 0.0001* and parameter values $\beta_1 = 0.9, \beta_2 = 0.999$. The batch size is set to 8. The total number of trainable parameters for the entire network, according to model summary is approximately 42.8M parameters. The training process is completed with 300 iterations.

6.6.2 Dataset for Monocular Depth Estimation

NYU Depth v2 is a dataset that provides images and depth maps for different indoor scenes captured at a resolution of 640 × 480 help of Kinect Camera. Even if the dataset contains 120K training samples and 654 testing samples [40], the method uses a 40K subset. Missing depth values are completed using the inpainting method which is proposed in [49]. The bound depth maps for this dataset is approximately 10 meters . The network produces predictions at half the input resolution, i.e. a resolution of 320 × 240.In training step, input images are taken at their original resolution and downsample the ground truth depths to 320 × 240. With help of the preprocessing steps which are explained previous chapter, cropping any of the input image-depth map pairs even if they contain missing pixels are not required. For test process, the depth map prediction of the test image after that, the model upsamples it by 2× to match the ground truth resolution and evaluate. At test time, the final output is calculated by taking the average of prediction of the image and the prediction of its mirror image. the Figure 6.3 illustrates the sample format of NYU v2 Dataset.

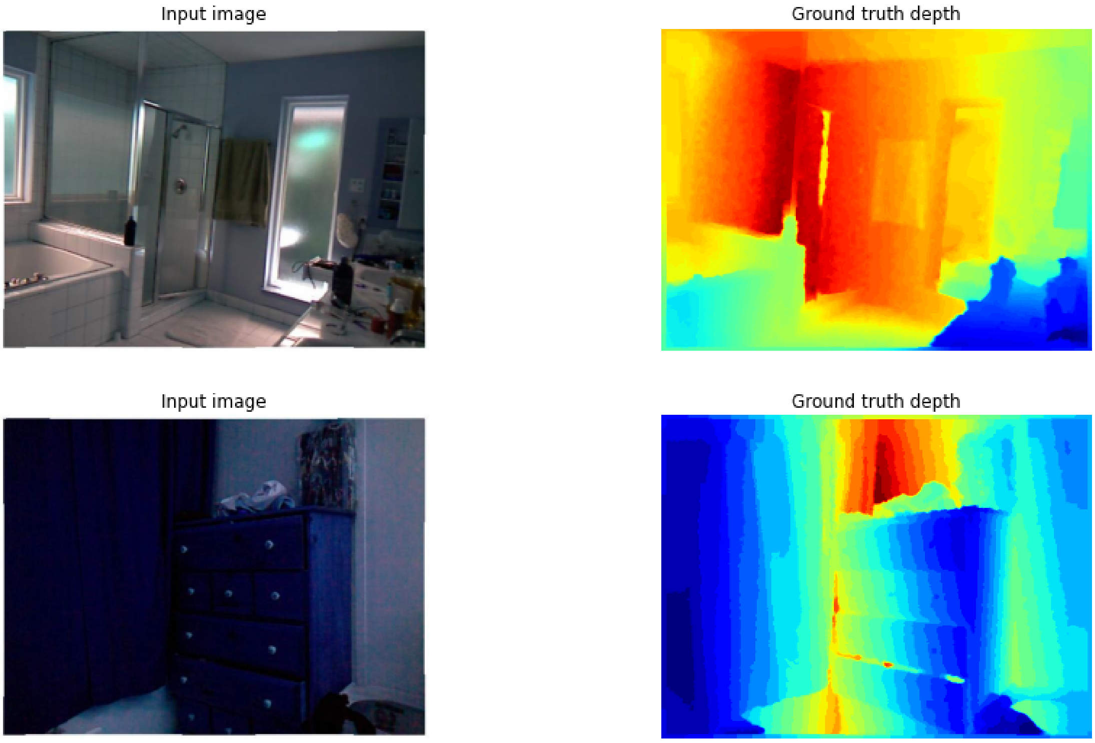


Figure 6.3: Sample Input Image for U-net Architecture

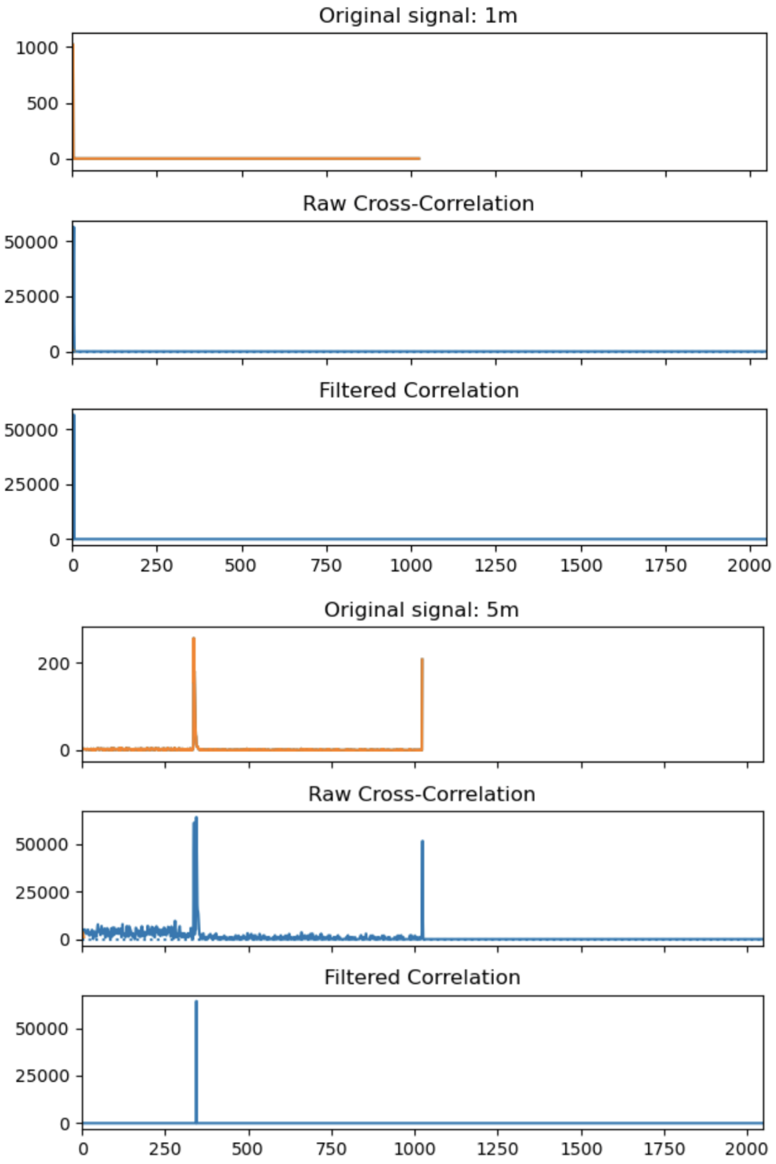
7
RESULTS

This part includes the main details of datasets which are exploited for the implementation and analysis, the results and error illustrations with help of sample input-output and graphs. After that, the comparison among these depth estimation approaches will be discussed.

In order to propose the results of the methods which are introduced in the previous sections, a Python program has been developed which is separated to 2 parts for ToF based distance estimation and U-net based depth map estimation. In ToF part, this software allows us handle the distance histograms to provide a result for estimation. After the MATLAB simulation of ToF sensor, threshold is applied to reduce the dimension and then deep regression is implemented. In the second part, in order to provide an implementation of image based depth estimation, sub-dataset of NYU Dataset V.2 is exploited (40k Images).

7.1 Results of ToF Based Distance Estimation

Time of Flight Camera simulation provides us histograms with synthetic data, after the the base distance assumption which is explained in the Chapter 4, the first data shows in Fig1, different peak points for every time stamp, make the required data space larger. In this case, in order to reduce the data, raw cross correlation is applied. The figure is given the below illustrates the raw correlation between destination signal (y value) and base signal (distance=0), after that threshold is applied and smooth correlation is obtained as new datum for the model. With help of this approach, observing difference between target and base becomes easier. Some wrong peak points can be seen after the correlation operation (e.g 5m distance) due to last time stamp value or some environmental conditions (e.g ambient light).



After creation of model, back propagation updates the weights according to loss value and then optimization step applies *L1* regularization. The graphs are based on standard deviation for every integer distance sample. To measure the sensitivity of the model, all distance estimations are calculated and then the average distance estimation for a destination distance are used for standard deviation.

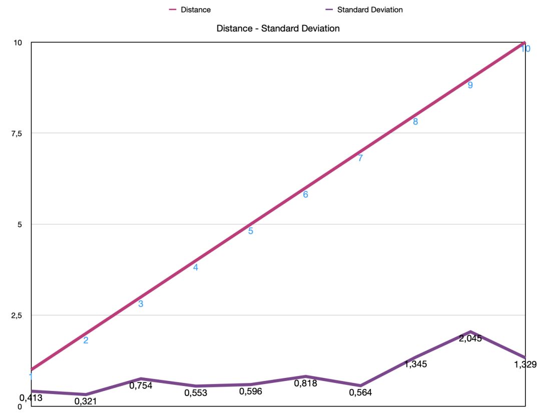


Figure 7.1: Single Layer, Neuron Count: 400, Accuracy: 70%

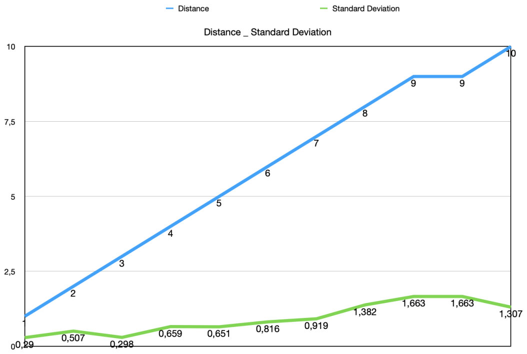


Figure 7.2: Multi Layer, Neuron Count: 400, Accuracy: 61%

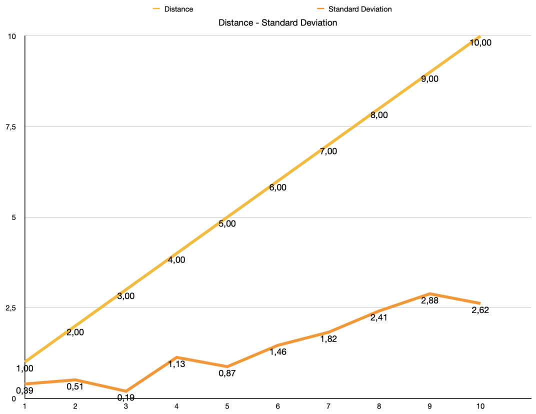


Figure 7.3: Multi Layer, Neuron Count: 400, Accuracy: 87%

Standard Deviation			
Distance (m)	Single Layer	Multi Layer: 3	Multi Layer: 5
1	0.413	0.290	0.393
2	0.321	0.507	0.516
3	0.754	0.298	0.191
4	0.553	0.659	1.130
5	0.596	0.651	0.871
6	0.818	0.816	1.465
7	0.564	0.919	1.826
8	1.345	1.382	2.415
9	2.045	1.663	2.880
10	1.329	1.307	2.622

The table shows the standard deviation for every integer distance in the dataset.

Summary, as one of the most important case for time-of-flight based distance estimation, when the destination distance is getting larger, sensitivity for estimation increases. This situation affects the accuracy negatively for longer distances. The model which provides us the most accurate results is multi layer deep regression algorithm. To make an effective comparison among the layers, all hyper-parameters such as learning rate, number of neurons, epoch number are selected same. According to Figure 7.3 the best accuracy is given by multi layer, even if the highest standard deviation. On the other hand, the lowest standard deviation is provided by single layer deep regression. There is trade-off between accuracy and standard deviation. Despite of higher accuracy, error ratio for this strategy is the highest.

7.2 Disparity Maps with Neural Network

To improve the performance of the disparity maps, Neural Networks are exploited. Completing of missing points and some benefits for exceptional textureless surfaces or conditions (e.g further objects, optical reflection and losing some details about corners and edges) affect the performance directly. As it is illustrated in Figure 7.6, implementation of NN for disparity maps provides us clearer results. Specifically, instead of disparity maps (Figure 7.5), reliability for edges is better in Figure 7.6. As a loss evaluation, the academic study [28] which is followed uses EPE (end-point-end). It is based on average error between estimated disparity and ground truth for every pixel. The images are taken from KITTI Dataset 2015, it is the one of the most common dataset for the approach and in this part a sample image is placed to demonstrate the positive effect of the Neural Networks.

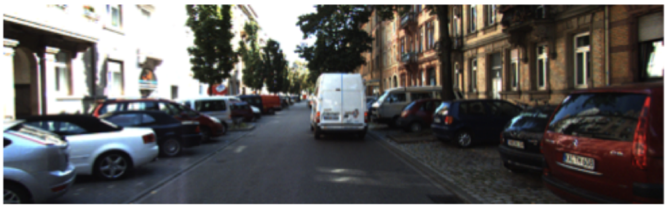


Figure 7.4: Original Image

Figure 7.5: Disparity Map

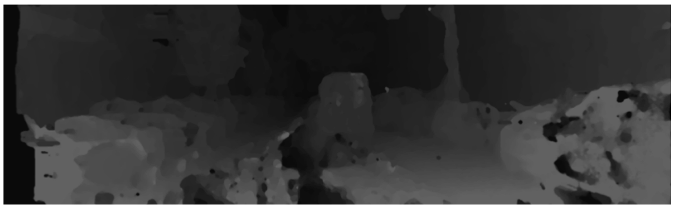


Figure 7.6: Estimated Disparity of Neural Network

7.3 Results of U-net

The output images are obtained from U- Architecture for monocular depth estimation. The dataset is splitted to train set, validation set and test set. The trainset contains 80% of all images and validation-set and test-set contain 10% of the image dataset.

Supervised learning is the one of the most common option for depth maps, U-net provides us more sensitive results to handle the relation among pixels. Even if the stereo devices provide better information a scene, single image sources are common in the real world. For monocular depth estimation, U-net model exploit 2x2 upsampling size and for convolutional layers, it uses 3x3 kernel size. The hyper parameters such as learning rate: 10^{-4} and epoch limit is : *10* to prevent the overfitting case. The best score end of the last epoch is 93% for training set , 91% for validation set and 83% for test set. Higher learning rates may lead to divergent error for the model, on the other hand, smaller learning rates can involve the overfitting. In order to improve the accuracy, k fold cross validation is integrated to span the dataset properly.

As it is illustrated with sample output images, some missing details are occurred for edges

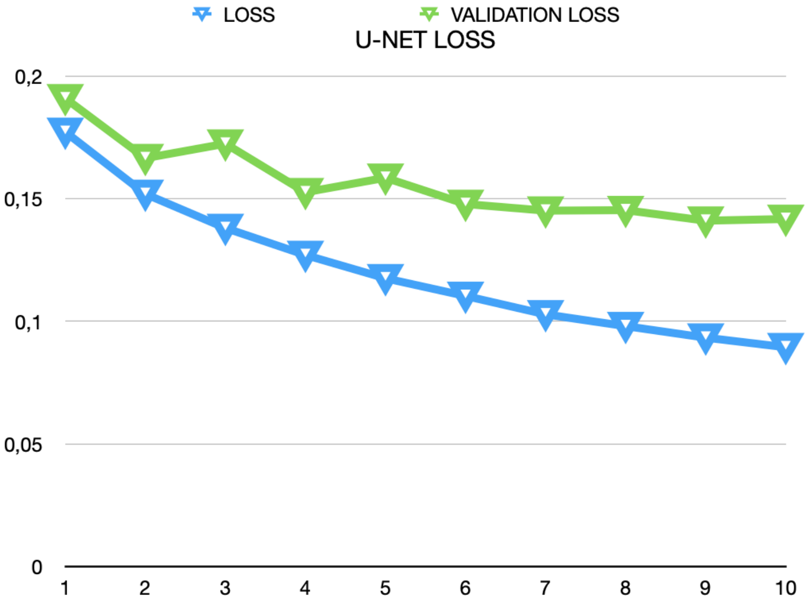


Figure 7.7: Loss Plot for U-net

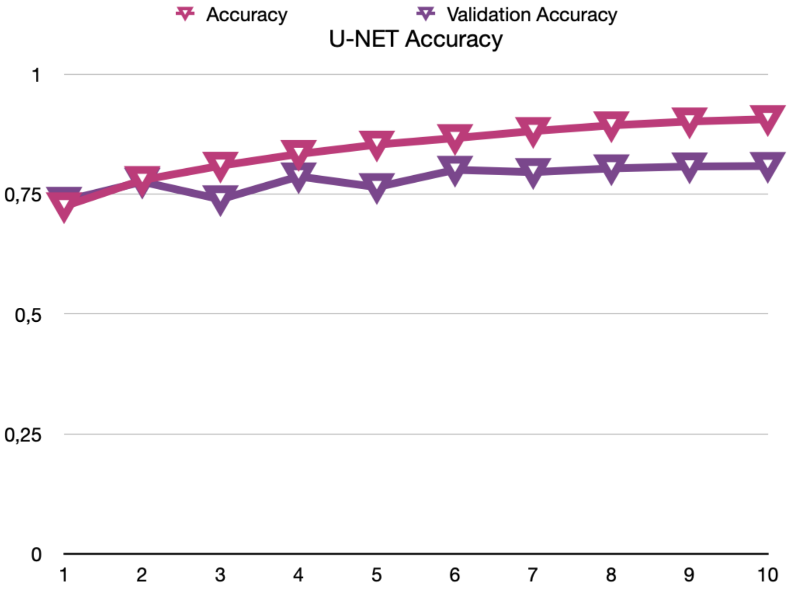


Figure 7.8: Accuracy

and corner information in the scenes. Increasing epoch number does not affect this drawback properly. At the same time, it reduces run-time performance. Instead of epoch limit, early stopping method can be integrated to the model.

Finally, the model produces the depth estimation with input image and its Ground Truth. As it is illustrated in below. The first column shows input images, the second column shows its Ground Map and then the third column shows the prediction.

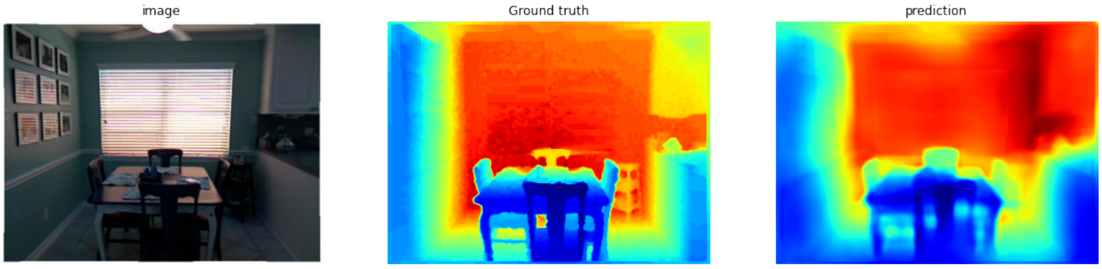


Figure 7.9

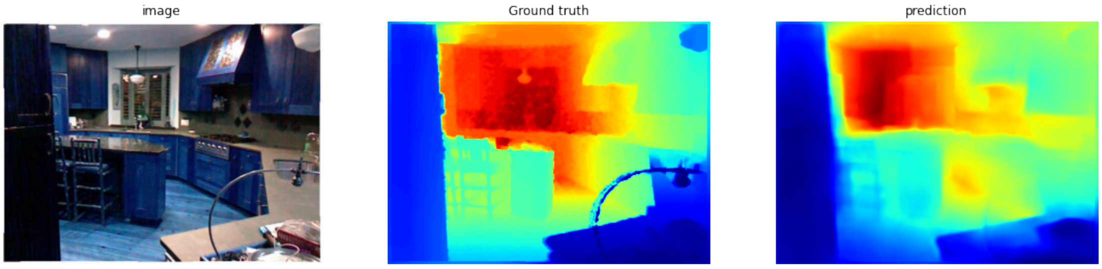


Figure 7.10

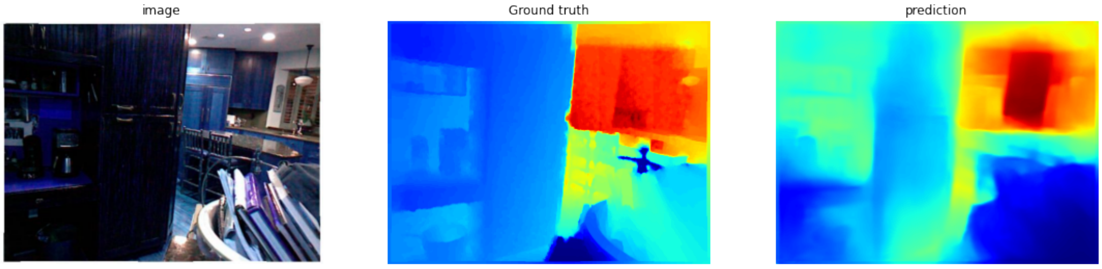


Figure 7.11

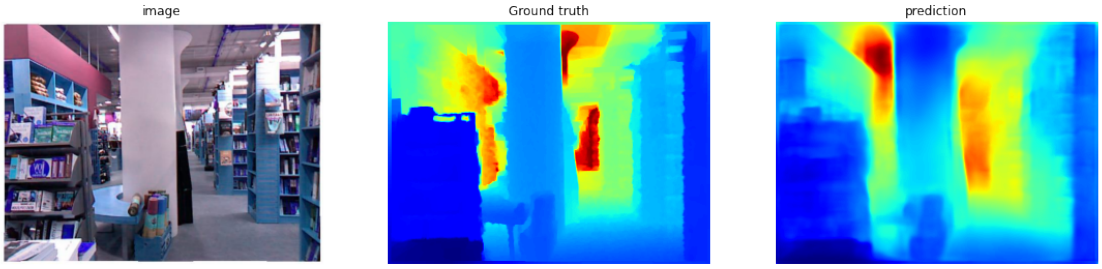


Figure 7.12

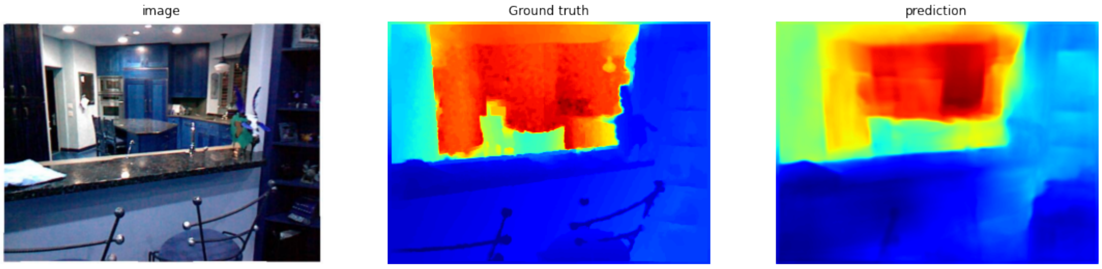


Figure 7.13

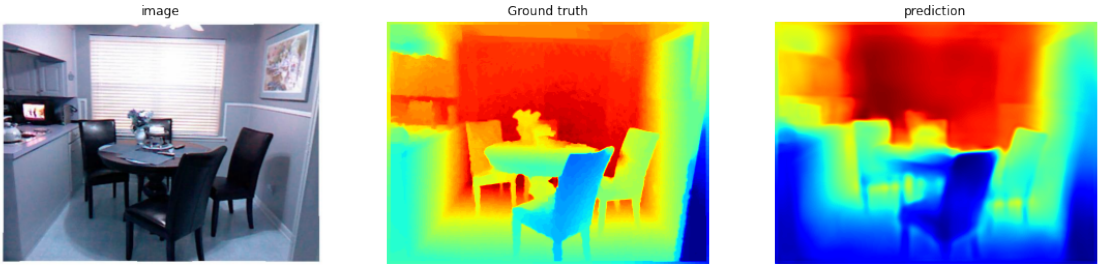


Figure 7.14

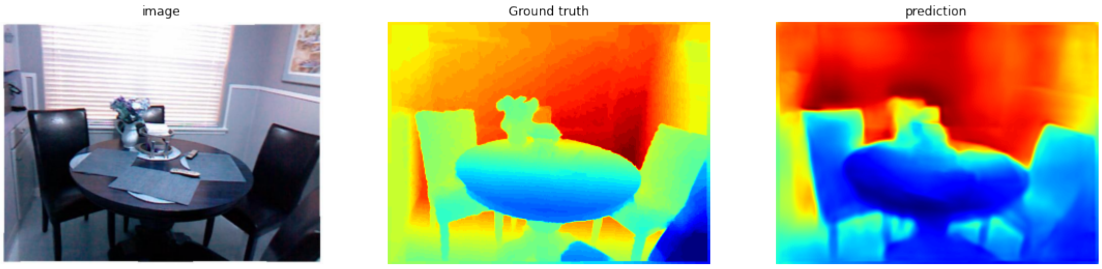


Figure 7.15

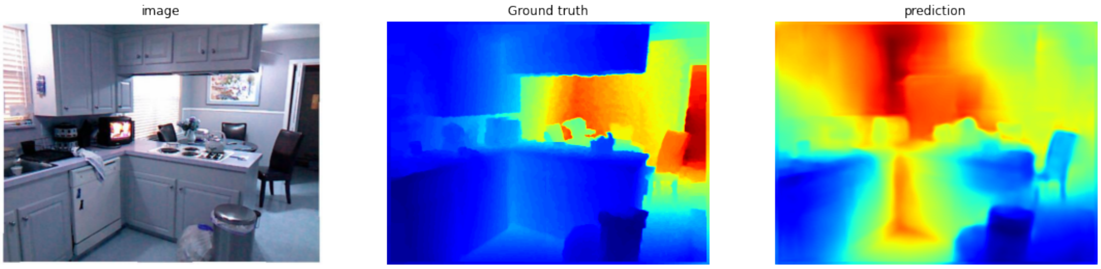


Figure 7.16

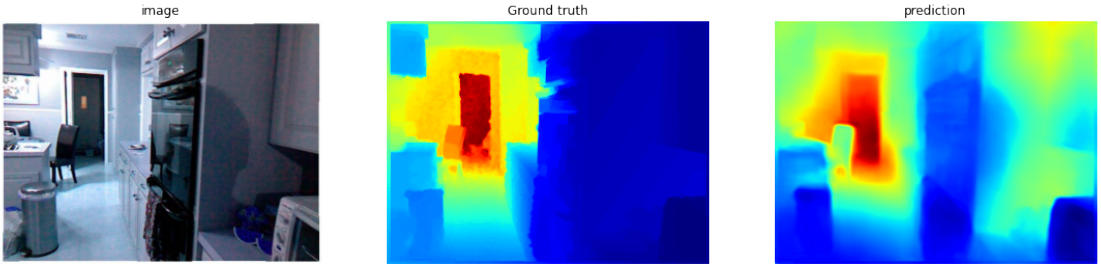


Figure 7.17

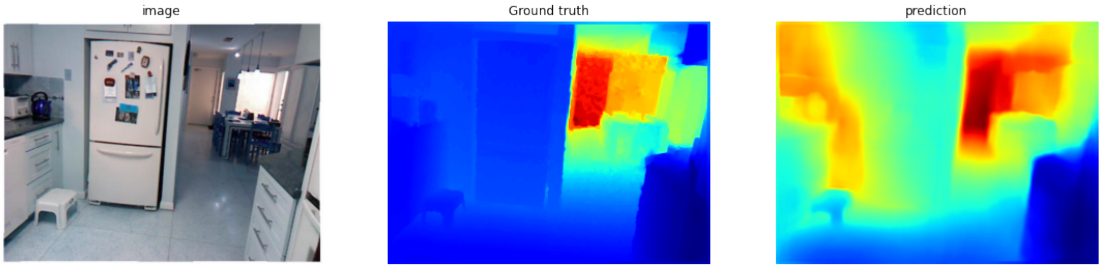


Figure 7.18

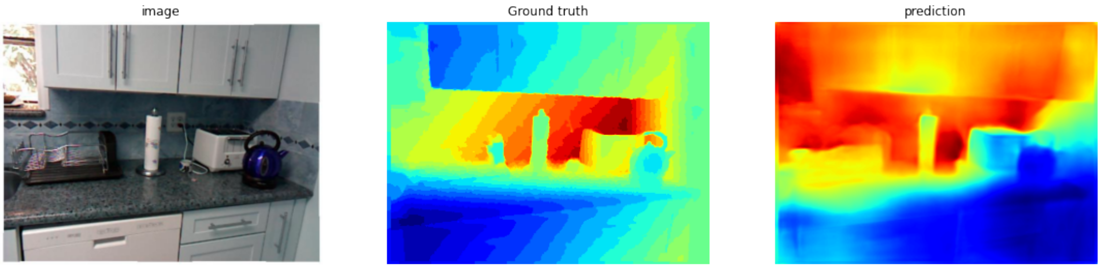


Figure 7.19

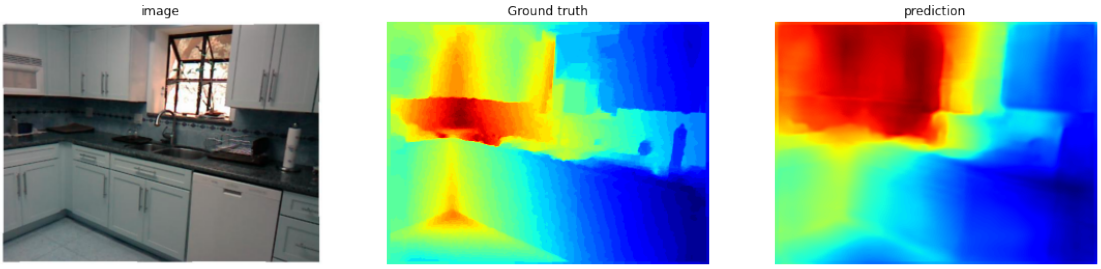


Figure 7.20

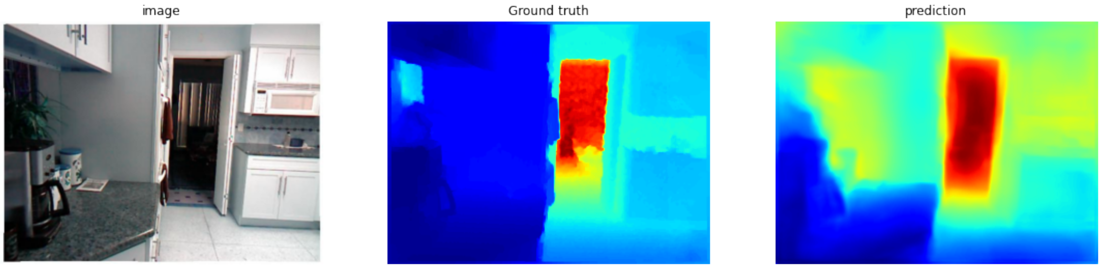


Figure 7.21

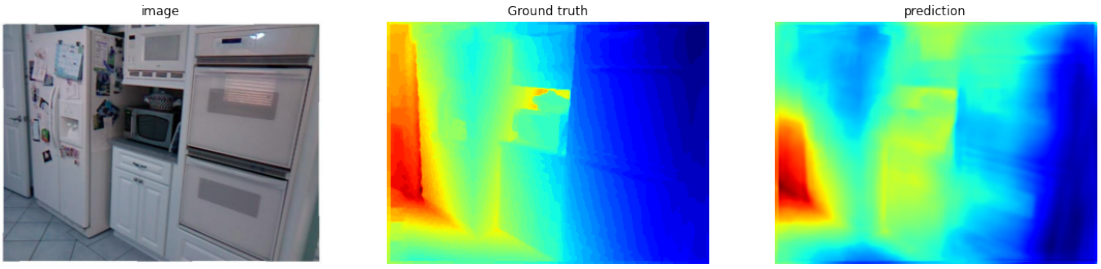


Figure 7.22

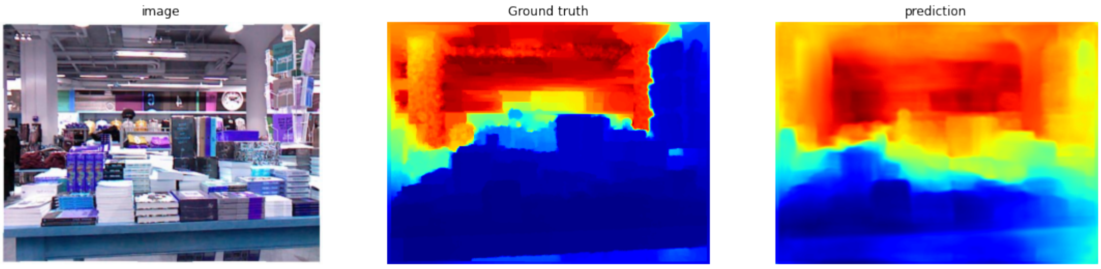


Figure 7.23

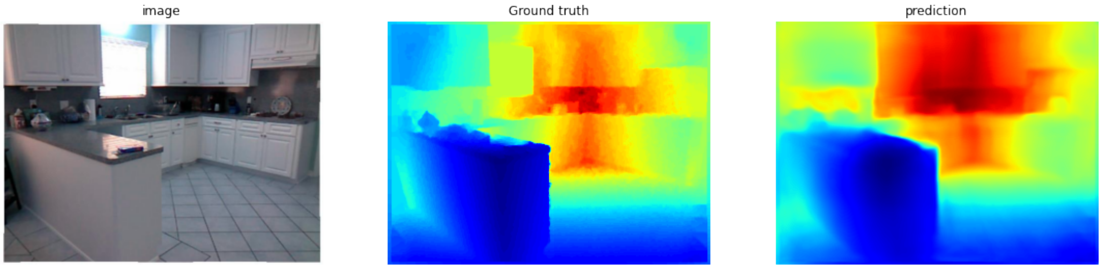


Figure 7.24

8
Conclusion

This thesis introduces 3D Data Estimation with help of the most common approaches in Computer Vision Field. Through, the first part provides a distance estimation tool via Single and Multiple Layer Deep Regression, stereo vision part demonstrates the advantages of Neural Networks for disparity maps and then last part implements U-net algorithm as a solution for monocular depth estimation.

In Time-of-Flight part shows the catalysed performance for distance estimation with help of Machine Learning Modelling which are supported by traditional threshold approach. The most common drawback for this task could be eliminated via data reduction, and then Regression Algorithm is implemented to satisfy to the expectations about real-time running performance. For stereo vision devices part demonstrates the effective contribution of Neural Networks to improve the disparity maps for depth estimation. Instead of additional calculation with more traditional approaches, Neural Networks could handle the situation properly. Monocular Depth estimation provides us robust results with an acceptable loss in edges and corners. Even if, only one RGBD image for a scene leads to many advantages in Computer Vision. In order to make a comparison among these methods, the results are observed to discuss about their drawbacks and benefits.

For an extension to improve the results, some improvements could be applied for data reduction and increase the sensor sensitivity to minimize to pile-up effect in time-of-flight part. As a future work for stereo devices, some pre-processing steps can be integrated to improve image quality and diminish the running time. Furthermore, some pre-processing steps can be considered for monocular depth estimation with U-net to preserve the edges and corners.

From widely diverse industries to daily life, time-of-flight sensors, stereo image devices and monocular devices are exploited for many tasks. Thus, we could be sure that in the future there are many developments and research will be released to provide better performance and scientific results.

References

[1] Matteo Perenzoni, Daniele Perenzoni and David Stoppa, "A 64×64-Pixels Digital Silicon Photomultiplier Direct TOF Sensor With 100-MPhotons/s/pixel Background Rejection and Imaging/Altimeter Mode With 0.14% Precision Up To 6 km for Spacecraft Navigation and Landing", IEEE Journal of Solid-state Circuits, Vol. 52, No. 1, January 2017.

[2] Ruibin Feng, David Rundle and Ge Wang, “Neural-Networks-Based Photon-Counting Data Correction: Pulse Pileup Effect”, April 2018.

[3] Zhe Wang, Hongsheng Li, Wanli Ouyang and Xiaogang Wang, "Learnable Histogram: Statistical Context Features for Deep Neural Networks", Dept. of Electronic Engineering, The Chinese University of Hong Kong, October 2018.

[4] C. Oprea, I. Pirnog, I. Marcu and M. Udrea, “Robust Pose Estimation Using Time-of-Flight Imaging”, Department of Telecommunications, Faculty of Electronics, Telecommunications and Information Technology, University Politehnica of Bucharest, Romania, October 2019.

[5] Gianluca Agresti, Ludovico Minto, Giulio Marin and Pietro Zanuttigh, "Deep Learning for Confidence Information in Stereo and ToF Data Fusion”, University of Padova, October 2017.

[6] Stefan May, David Droeschel, Dirk Holz and Christoph Wiesen, “3D Pose Estimation and Mapping with Time-of-Flight Cameras”, Fraunhofer Institute for Intelligent Analysis and Information Systems (IAIS) Schloss Birlinghoven, Stefan Fuchs German Aerospace Center (DLR) Institute of Robotics and Mechatronics, January 2008.

[7] Rene Ranftl, Katrin Lasinger, David Hafner, Konrad Schindler and Vladlen Koltun, “Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-shot Cross-Dataset Transfer”, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. XX, No. XX, August 2020.

[8] Yang Yang, “Attention-based Dual Supervised Decoder for RGBD Semantic Segmentation”,Hubei University of Technology , 2022 .

[9] Shuochen Su, Felix Heide, Gordon Wetzstein and Wolfgang Heidrich, “Deep End-to-End Time-of-Flight Imaging”, 2018.

[10] C. Dal Mutto, P. Zanuttigh, S. Mattoccia and G. Cortelazzo. “Locally Consistent ToF and Stereo Data Fusion”, In: Proceedings of the 12th international conference on Computer Vision - Volume Part I. ECCV’12. Florence, Italy:Springer-Verlag, 2012, pp. 598–607 (cit. on pp. 3, 17, 19).

[11] Giulio Marin, "Confidence Estimation of ToF and Stereo Data for 3D Data Fusion", University of Padova, 2013.

[12] Michael P. Sheehan, Julián Tachella and Mike E. Davies, "A Sketching Framework for Reduced Data Transfer in Photon Counting Lidar", Feb 2021.

[13] R. K. Henderson, N. Johnston, H. Chen, D. D. Li, G. Hungerford, R. Hirsch, D. McLoskey, P. Yip, and D. J. S. Birch, "A 192×128 time correlated single photon counting imager in 40nm CMOS technology," in ESSCIRC 2018 - IEEE 44th European Solid State Circuits Conference (ESSCIRC), 2018, pp. 54–57.

[14] F. M. Della Rocca, H. Mai, S. W. Hutchings, T. Al Abbas, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "A 128 × 128 SPAD dynamic vision-triggered time of flight imager," in ESSCIRC 2019 - IEEE 45th European Solid State Circuits Conference (ESSCIRC), 2019, pp. 93–96.

[15] F. Mattioli Della Rocca, H. Mai, S. W. Hutchings, T. A. Abbas, K. Buckbee, A. Tsiamis, P. Lomax, I. Gyongy, N. A. W. Dutton, and R. K. Henderson, "A 128 × 128 SPAD motion-triggered time-offlight image sensor with in-pixel histogram and column-parallel vision processor," IEEE Journal of Solid-State Circuits, vol. 55, no. 7, pp. 1762–1775, 2020.

[16] C. Zhang, S. Lindner, I. M. Antolovic, J. Mata Pavia, M. Wolf, and E. Charbon, "A 30-frames/s, 252 × 144 SPAD flash lidar with 1728 dual-clock 48.8-ps tdcs, and pixel-wise integrated histogramming," IEEE Journal of Solid-State Circuits, vol. 54, no. 4, pp. 1137–1151, 2019.

[17] A. Kadambi and P. T. Boufounos, "Coded aperture compressive 3-D lidar," in 2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2015, pp. 1166–1170.

[18] A. Halimi, P. Ciuciu, A. Mccarthy, S. Mclaughlin, and G. S. Buller, "Fast adaptive scene sampling for single-photon 3D lidar images," in IEEE CAMSAP 2019 - International Workshop on Computational Advances in Multi-Sensor Adaptive Processing, Le Gosier (Guadeloupe), France, Dec. 2019.

[19] Y. Altmann and S. McLaughlin, "Range estimation from single-photon lidar data using a stochastic em approach," in 2018 26th European Signal Processing Conference (EUSIPCO), 2018, pp. 1112–1116.

[20] L. P. Hansen, "Large sample properties of generalized method of moments estimators," Econometrica, vol. 50, no. 4, pp. 1029–1054, 1982.

[21] A. Hall, Generalized Method of Moments, 11 2007, pp. 230 – 255.

[22] R. Gribonval, G. Blanchard, N. Keriven, and Y. Traonmilin, “Statistical learning guarantees for compressive clustering and compressive mixture modeling,” arXiv preprint arXiv:2004.08085, 2020.

[23] N. Keriven, A. Bourrier, R. Gribonval, and P. Pérez, “Sketching for largescale learning of mixture models,” Information and Inference: A Journal of the IMA, vol. 7, no. 3, pp. 447–508, 2018.

[24] K. He, J. Sun, and X. Tang. “Guided Image Filtering”. In: Computer Vision – ECCV 2010. Ed. by K. Daniilidis, P. Maragos, and N. Paragios. Springer Berlin Heidelberg, 2010, pp. 1–14 (cit. on p. 28).

[25] A. Hosni, C. Rhemann, M. Bleyer, C. Rother, and M. Gelautz. “Fast CostVolume Filtering for Visual Correspondence and Beyond”. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) 35.2 (2013),pp. 504–511 (cit. on p. 28).

[26] OpenCV. url: http://www.opencv.org (cit. on p. 32).

[27] D. Scharstein, R. Szeliski, and R. Zabih. “A taxonomy and evaluation of dense two-frame stereo correspondence algorithms”. In: Stereo and Multi-Baseline Vision, 2001. (SMBV 2001). Proceedings. IEEE Workshop on. 2001, pp. 131–140 (cit. on pp. 28, 32).

[28] Wen-Nung Lie, Hung-Ta Chiu, and Jui-Chiu Chiang, “Disparity Map Estimation From Stereo Image Pair Using Deep Convolutional Network”, Department of Electrical Engineering, Center for Innovative Research on Aging Society (CIRAS), Advanced Institute of Manufacturing with High-tech Innovations (AIM-HI),National Chung Cheng University (CCU), Taiwan, 2020.

[29] Z. Zhou, M. Siddiquee, N. Tajbakhsh, and J. Liang, “UNet++: A nested u-net architecture for medical image segmentation”, Proc. of Deep Learning in Medical Image Analysis and Multimodal Learning for Clinical Decision Support, pp.3-11, 2018.

[30] A. Kendall, H. Martirosyan, S. Dasgupta, P. Henry, R. Kennedy, A. Bachrach, and A. Bry, “End-to-end learning of geometry and context for deep stereo regression,” Proc. of IEEE Int’l Conf. on Computer Vision, 2017.

[31] Richard Chen, Faisal Mahmood, Alan Yuille, and Nicholas J. Durr, "High Quality Monocular Depth Estimation via Transfer Learning", March 2019.

[32] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. In Proceedings of the 4th International Conference on 3D Vision (3DV 2016), pages 239–248. IEEE, 2016.

[33] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe. Monocular depth estimation using multi-scale continuous crfs as sequential deep networks. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[34] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci. Structured attention guided convolutional neural fields for monocular depth estimation. In Proceedings of the 2018 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[35] J. Jiao, Y. Cao, Y. Song, and R. Lau. Look deeper into depth: Monocular depth estimation with semantic booster and attention-driven loss. In Proceedings of the 2018 IEEE European Conference on Computer Vision (ECCV), September 2018.

[36] A. Odena, C. Olah, and J. Shlens. Conditional image synthesis with auxiliary classifier gans. In Proceedings of the 34th International Conference on Machine Learning (PLMR 2017), Proceedings of Machine Learning Research, pages 2642–2651, 2017.

[37] M. Drozdzal, E. Vorontsov, G. Chartrand, S. Kadoury, and C. Pal. The Importance of Skip Connections in Biomedical Image Segmentation. pages 179–187, 2016.

[38] S. Bianco, R. Cadene, L. Celona, and P. Napoletano. Bench-mark analysis of representative deep neural network architectures. IEEE Access, 6:64270–64277, 2018.

[39] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition, pages 248–255, 2009.

[40] D. Eigen, C. Puhrsch, and R. Fergus. Depth map prediction from a single image using a multi-scale deep network. In NIPS, 2014.

[41] H. Fu, M. Gong, C. Wang, N. Batmanghelich, and D. Tao. Deep ordinal regression network for monocular depth estimation. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2002–2011, 2018.

[42] C. Godard, O. M. Aodha, and G. J. Brostow. Unsupervised monocular depth estimation with left-right consistency. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 6602–6611, 2017.

[43] Z. Hao, Y. Li, S. You, and F. Lu. Detail preserving depth estimation from a single image using attention guided networks. 2018 International Conference on 3D Vision (3DV), pages 304–313, 2018.

[44] G. Huang, Z. Liu, L. van der Maaten, and K. Q. Weinberger. Densely connected convolutional networks. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 2261–2269, 2017.

[45] P.-H. Huang, K. Matzen, J. Kopf, N. Ahuja, and J.-B. Huang. Deepmvs: Learning multi-view stereopsis. 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 2821–2830, 2018

[46] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[47] I. Laina, C. Rupprecht, V. Belagiannis, F. Tombari, and N. Navab. Deeper depth prediction with fully convolutional residual networks. 2016 Fourth International Conference on 3D Vision (3DV), pages 239–248, 2016.

[48] J. Lehtinen, J. Munkberg, J. Hasselgren, S. Laine, T. Karras, M. Aittala, and T. Aila. Noise2Noise: Learning image restoration without clean data. In J. Dy and A. Krause, editors, Proceedings of the 35th International Conference on Machine Learning, volume 80 of Proceedings of Machine Learning Research, pages 2965–2974, Stockholmsmssan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[49] A. Levin, D. Lischinski, and Y. Weiss. Colorization using optimization. ACM Trans. Graph., 23:689–694, 2004.

[50] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg, A. Dosovitskiy, and T. Brox. Demon: Depth and motion network for learning monocular stereo. 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pages 5622–5631, 2017.

[51] J. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli, Image quality assessment: from error visibility to structural similarity, IEEE Transactions on Image Processing, 13:600–612, 2004.

[52] Ibraheem Alhashim KAUST and Peter Wonka KAUST, "High Quality Monocular Depth Estimation via Transfer Learning", March 2019.

[53] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei., "Imagenet: A large-scale hierarchical image database. 2009 IEEE Conference on Computer Vision and Pattern Recognition", 2009.