МІНІСТЕРСТВО ОСВІТИ І НАУКИ УКРАЇНИ

НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ

Кафедра авіаційних комп’ютерно-інтегрованих комплексів

ДОПУСТИТИ ДО ЗАХИСТУ

Завідувач кафедри

Синєглазов В.М.

“ ____ ” _____2021.

ДИПЛОМНАРОБОТА

(ПОЯСНЮВАЛЬНА ЗАПИСКА)

ВИПУСНИКА ОСВІТНЬО-КВАЛІФІКАЦІЙНОГО РІВНЯ

“БАКАЛАВР”

Тема: Система оцінки глибини зображення за потоковим відео

Виконав: **Сич О. С.**

Керівник:**к.т.н.** **Василенко М. П.**

Нормоконтролер:к.т.н. **Туріцин М. Ф.**

Київ – 2021

EDUCATION AND SCIENCE MINISTRY OF UKRAINE

NATIONAL AVIATION UNIVERSITY

COMPUTER-INTEGRATED COMPLEXES DEPARTMENT

ADMIT TO DEFENSE

Head of department

V. M. Sineglazov

“____” _____ 2021.

BACHELOR WORK

(EXPLANATORY NOTES)

Topic: Image depth estimation system by streaming video

Done by: **Sych O. S.**

Supervised by: **Vasylenko M. P.**

Normcontrolled by: **Tupitsyn M. F.**

Kyiv 2021

НАЦІОНАЛЬНИЙ АВІАЦІЙНИЙ УНІВЕРСИТЕТ

Факультет аеронавігації, електроніки та телекомунікацій

Кафедра авіаційних комп’ютерно-інтегрованих комплексів

Освітній ступінь бакалавр

Спеціальність: 151 " Автоматизація та комп'ютерно-інтегровані технології"

ЗАТВЕРДЖУЮ

Завідувач кафедри

Синєглазов В.М.

“ ____ ” _____2021 р.

ЗАВДАННЯ

на виконання дипломної роботи студента

Сича Олексія Сергійовича

1. **Тема проекту (роботи):** “ Система оцінки глибини зображення за потоковим відео ”
2. **Термін виконання проекту (роботи):** з 10.05.2021 р. до 11.06.2021 р.
3. **Вихідні данні до проекту (роботи):** оптичний метод визначення дальності до об'єкта, карти глибини, алгоритми калібрування камер, середовище Matlab.
4. **Зміст пояснювальної записки (перелік питань, що підлягають розробці):** 1.Актуальність системи оцінки глибини зображення за потоковим відео; 2. Огляд існуючих методів; 3. Огляд теоретичної інформації з приводу рішення задачі карти глибини; 4. Розробка системи оцінки глибини зображення за потоковим відео.
5. **Перелік обов’язкового графічного матеріалу:** 1. Блок-схема калібрування камери; 2. Положення камер у просторі; 3. Структурна схема системи роботи з одним зображенням; 4. Карта розбіжностей у кольорі; 5. Таблиця перевірки точності щодо реального та виміряного діапазону за допомогою карти глибини; 6. Структурна схема пристрою виведення потокової 3D-сцени.

6. Календарний план-графік

№ пор.	Завдання	Термін виконання	Відмітка про виконання
1.	Отримання завдання	10.05.2021 – 11.05.2021	
2.	Формування мети та основних завдань дослідження	12.05.2021 – 13.05.2021	
3.	Аналіз існуючих методів	14.05.2021 – 19.05.2021	
4.	Теоретичний розгляд рішення задачі	20.05.2021 – 25.05.2021	
5.	Розробка структури системи оцінки глибини зображення	25.05.2021 – 30.05.2021	
6.	Розробка програмного та апаратного забезпечення системи оцінки глибини зображення за потоковим відео	30.05.2021 – 05.06.2021	
7.	Оформлення пояснювальної записки	05.06.2021 – 07.06.2021	
8.	Підготовка презентації та роздаткового матеріалу	08.06.2021 – 11.06.2021	

7. Дата видачі завдання: “10” травня 2021 р.

Керівник дипломної роботи _____ Василенко М. П.

(підпис керівника) (П.І.Б.)

Завдання прийняла до виконання _____ Сич О. С.

(підпис випускника) (П.І.Б.)

NATIONAL AVIATION UNIVERSITY

Faculty of aeronavigation, electronics and telecommunications

Department of Aviation Computer Integrated Complexes

Educational level bachelor

Specialty: 151 "Automation and computer-integrated technologies"

APPROVED

Head of Department

Sineglazov V. M.

"____" _____2021

TASK

For the student's thesis

Sych Oleksii Serhiyovych

1. **Theme of the project:** " Image depth estimation system by streaming video "
2. **The term of the project (work):** from May 10, 2021 until June 11, 2021
3. **Output data to the project (work):** optical method for determining the distance to the object, depth maps, camera calibration algorithms, Matlab environment.
4. **Contents of the explanatory note (list of questions to be developed):**

1. Relevance of the image depth estimation system for streaming video 2. Review of existing methods; 3. Review of theoretical information about the solution of the depth map problem; 4. Development of an image depth estimation system based on streaming video.

5. **List of compulsory graphic material:**

1. Block diagram of the camera calibration; 2. Position of cameras in space; 2. Block diagram of the system work with one image; 3. Disparity Map in color; 4. Accuracy check table for real and measured range using depth map; 5. Block diagram of the device for outputting a streaming 3D scene.

6. **Planned schedule:**

№	Task	Execution term	Execution mark
1.	Task	10.05.2021 – 11.05.2021	
2.	Purpose formation and describing the main research tasks	12.05.2021 – 13.05.2021	
3.	Analysis of existing methods	14.05.2021 – 19.05.2021	
4.	Analysis of existing systems	20.05.2021 – 25.05.2021	
5.	Development of the structure of the image depth estimation system	25.05.2021 – 30.05.2021	
6.	Development of software and hardware for image depth estimation system based on video streaming	30.05.2021 – 05.06.2021	
7.	Making an explanatory note	05.06.2021 – 07.06.2021	
8.	Preparation of presentation and handouts	08.06.2021 – 11.06.2021	

7. **Date of task receiving:** "10" May 2021

Diploma thesis supervisor Vasylenko M. P.

(signature)

Issued task accepted Sych O. S.

(signature)

РЕФЕРАТ

Пояснювальна записка до дипломної роботи «Система оцінки глибини зображення за потоковим відео»: 90 с., 39 рис., 2 табл., 23 літературних джерела.

Об’єкт дослідження: процес оцінки глибини зображення.

Мета роботи: вдосконалення методів оцінки глибини зображення та розробки нової системи, що здатна з порівняно невеликими обчислювальними затратами здійснювати таку оцінку в реальному часі.

Для досягнення цієї мети необхідно розв’язати наступні завдання:

- проаналізувати існуючі методи оцінки глибини зображення;
- проаналізувати існуючі апаратні засоби, придатні для реалізації розглянутих методів;
- на основі проведеного аналізу здійснити вибір методу який би дозволив здійснювати оцінку глибини зображення з мінімальними обчислювальними та фінансовими затратами;
- розробити програмне та апаратне забезпечення реалізації обраного методу;
- провести експериментальне дослідження роботи розробленої системи.

Предмет дослідження: розробка методу оцінки глибини зображення на основі неперервного відео потоку.

Методи дослідження: теоретична фізика, теорія оптики, теорія обробки зображення та машинного зору.

СТЕРЕО ЗІР; КАРТА РОЗРІЗНОСТІ; КАРТА ГЛИБИНИ; КАЛІБРУВАННЯ; РЕКТИФІКАЦІЯ; 3D СЦЕНА.

ABSTRACT

Explanatory note to the thesis " Image depth estimation system by streaming video ": 90 p., 39 figures, 2 tables, 23 literary resources.

The object of research: image depth estimation process.

The purpose of the work: improving methods for estimating image depth and developing a new system that is able to perform such an assessment in real time with relatively low computational costs.

To achieve this purpose, it must be solved the following tasks:

- to analyze the existing methods for estimating image depth;
- to analyze the existing hardware suitable for the implementation of the considered methods;
- on the basis of the analysis to make a choice of a method which would allow to carry out an estimation of depth of the image with the minimum computing and financial expenses;
- to develop software and hardware for the implementation of the selected method;
- to conduct an experimental study of the developed system.

Subject of research: development of a method for estimating the depth of the image based on a continuous video stream.

Methods of research: theoretical physics, theory of optics, theory of image processing and machine vision.

STEREO VISION; DISPARITY MAP; DEPTH MAP; CALIBRATION; RECTIFICATION; 3D SCENE.

CONTENT

GLOSSARY

3D - Three-dimensional space

CV – Computer vision

AI – Artificial intelligence

VR – Virtual reality

AR – Augmented reality

CAD - Created architectural projects into stereo images

EXIF - Exchangeable image file format

GPS - Global Positioning System

RFID - Radio-frequency identification

SDS-TWR - Symmetric Double Sided Two Way Ranging

RTT - Round Trip Time

ACK – Acknowledgement

PC – Portable Computer

CMOS - Complementary-symmetry/metal-oxide semiconductor

LiDAR - Light Detection and Ranging

USGS - United States Geological Survey

NOAA - National Oceanographic and Atmospheric Administration

NASA - National Aeronautics and Space Administration

IMU - Inertial Measurement Unit

GNSS - Global navigation satellite system

NN – neural network

RADAR - Radio detection and ranging

RS – Radar Station

EMW - Electromagnetic waves

INTRODUCTION

Today, the tasks of computer vision are becoming very relevant, more and more people are automating work in production due to some kind of software processes and machine devices, which can make job easier or more accurate. Based on this, it was decided to consider in detail the problem of stereo vision without using neural networks, or other more complex methods, since their use required costly methods of training, setting and controlling parameters.

The main task was to create a mechanism taking into account the price and quality, due to the fact that there is no cheap analogue on the internet market, which was suitable for the task of simple recognition of 3D scenes and made it possible to analyze the environment in which it is located, namely, to find out at what distance objects are located, what is their size, and so on.

In the course of the work, the method of using two web cameras was chosen, which were configured and calibrated for the task of stereo vision. The conditions of projective geometry and the relationship between the two cameras are also considered, since without this, the operation of the main algorithm of the work could not be successful at all. An algorithm and program have been created for the device to operate in streaming mode, which allows directly know the exact characteristics in LIVE video mode.

CHAPTER 1. RELEVANCE OF THE WORK

The work is based on computer vision, this is a promising technology that every year becomes better and more interesting for use for any purpose, therefore, arises the question of the relevance not only of the image depth map, but also computer vision technologies at all [1].

To begin with, it is worth to consider what computer vision is. Computer vision is a field of artificial intelligence related to image and video analysis. It includes a set of techniques that empower the computer to "see" and extract information from what it sees. The systems consist of a photo or video camera and specialized software that identifies and classifies objects. Such technologies are able to analyze images (photos, pictures, videos, barcodes) as well as faces and emotions.

Machine learning technologies are used to teach a computer to "see". A lot of data is collected that allow to isolate features and combinations of features for further identification of similar objects.

What are the advantages of using in different areas [2]?

1) Security. Access control systems based on face recognition are applicable in almost all areas: from business centers and company offices to banks and restaurants.

2) Service. Due to quick face identification, it is possible to shorten the time of customer service and offer personalized services.

3) Strengthening human capabilities. Computer vision allows to see what a person may not notice. This is especially true in medicine (analysis of X-rays and other images) and industry (detection of defects).

4) Reducing the time for routine tasks. Recognition usually takes a few seconds. The person will be considered a shelf in the store for proper product display much longer.

5) Autonomy. The development of unmanned vehicles and robots is impossible without computer vision.

What is 3D Scene Modeling?

Back in 2009, David McKinnon of the Queensland University of Technology (Australia) developed a 3DSee program [3] that generates 3D models from 5-15 photographs. An important condition: all photos must have at least 80-90% overlap. It took McKinnon eight years to develop.

Creation of 3D scenes is in demand in construction, interior design, military affairs, animation. Hollywood is already using this technology to faithfully reproduce lighting, the placement of actors and sets - to save money on technically demanding filming. Manufacturers use such 3D models to train robots that need to move in space along a certain route and overcome obstacles. 3D scanners are suitable for identity authentication, virtual fitting of clothes, and many other things. Already now, using a smartphone, it can be shot a person from different angles and get a 3D avatar.

Victor Lempitsky, head of Samsung AI Center, professor at Skoltech, in his presentation at OpenTalks AI noted that 3D scene modeling was the focus of CV specialists in 2020. So far, it is difficult for neural networks to reproduce in detail some textures - such as tree foliage or hair - and create full-fledged 360 ° models. But in the near future, it will replace 3D designers and animators: it will be able to create renders of buildings and interiors, animated presentations and VR simulations of objects themselves. For example, Google's NeRF technology already generates realistic 3D images that are used to create AR and VR environments.

Computer vision in transport: using a drone.

Computer vision is a necessary component for the development of autonomous land, air and sea transport. Technology helps machines to be navigated in space. Face recognition systems are used to ensure security at transport infrastructure facilities: train stations, airports, metro stations. In the future, a person will also become a ticket for any type of passenger transport. However, this is not possible yet due to the current legislation: it is possible to register for a plane only with a passport.

Computer vision technologies are able to analyze the occupancy rate of parking lots, providing information on the optimization of the urban transport system.

As for stereo vision [4], people receive about 80 percent of all information about the world around them through the organs of vision. At the earliest stages of evolutionary development, the visual apparatus of animals acquired a pair of eyes and the ability to see the world in three dimensions. People can easily determine the size, shape and distance of objects, isolate the whole from many details and systematize images. For this options people and animals are helped by the analysis of perspective distortions of objects, the true dimensions of which are known in advance, the language of chiaroscuro, the effects of aerial perspective, color distortions and uneven displacement of objects during the movement of the observer.

The mechanisms listed above belong to the category of monocular spatial vision, which allow people to confidently orient ourselves in space even with one eye.

Thanks to monocular vision, each person is able to adequately assess the space in a photograph or a painting by an artist, but still it is not possible to get a sense of the real depth of space. Therefore, the most accurate and effective instrument for space perception is binocular vision (stereo vision).

When solving problems of transmission and processing of volumetric images, special attention should be paid to those in which the reproduced volumetric images must correspond to the optimal conditions for their perception by the human vision apparatus. For this, the process of human observation of volumetric objects was considered.

In this work, for a more detailed and deep analysis, the task of research is not based on the results of perception, but on the results of a subjective assessment of the program, which displays an illustration of the general position of bodies in space.

Subjective assessments are put down by the machine, based on the inherent algorithm and threshold values determined in the process of analysis and comparison of previously obtained ones with current assessments.

This experiment will make it possible to correct the results obtained by a person, more precisely, after comparing the two results, it will be possible to obtain an "ideal" (reference) set of values, which will be used to assess the stereo vision (binocular vision) of any subject.

Binocular vision gives an idea of the depth of the visible space and the volume of objects in it, with great perfection. However, not only thanks to binocular vision, a person gets a visual idea of the depth of the visible space and the volume of objects in it. Stereoscopic perception exists due to many psychophysiological factors of the visual process.

Images that provide a perception of space with the same degree of depth localization as in natural viewing of real space are called stereoscopic, spatial or volumetric images. Images that provide a slightly lower degree of localization of the depth of the depicted space are usually called relief or plastic. And finally, images with an even lower degree of localization of the depth of space are considered flat images.

A three-dimensional image of an object can be real (i.e., a physical copy of a real object) or an imaginary, optical image, similar to the visible image of objects

in a mirror. It is such spatial images, artificially recreated in the mind of the observer as a result of psychophysiological sensations, that are understood as stereoscopic images.

The important tasks are the study of pathologies of stereovision, fatigue of stereovision of a human operator. Deterioration of stereo vision leads to difficulties in assessing the distance of objects and their position in the surrounding space.

The presence of stereovision is necessary for operators of complex control systems, in a number of professions associated with particularly precise and delicate production operations, when working with binocular and stereoscopic devices.

The idea of a three-dimensional image stems from the principle of human vision, that is, the perception of objects with two eyes (binocular vision).

Currently, numerous studies are being carried out on places where, for various reasons, a person cannot work. These are, for example, the study of deep-sea depressions in the seas and oceans, studies at sufficiently low temperatures, intolerable by humans, and much more. In such conditions, scientists come to the aid of numerous equipment, in particular, robotics, which performs certain assigned commands and actions and transmits to the operator all the information received: measured characteristics and parameters, pictures from the work site, as well as video information so that the operator has the opportunity to monitor the performance works in real time.

It is desirable that the information provided to the operator is displayed in a form that is familiar and convenient for him. To improve the operator's orientation at the work site, it is proposed to use a device that allows the formation of a three-dimensional image. This will make it possible to more accurately determine the position of objects relative to each other, as well as increase the likelihood of correct identification of objects and the range of visibility compared to monocular vision.

If turn to the history of the emergence of stereo, then the desire of a person to obtain a three-dimensional image from flat pictures can be found back in the 15th century in the works of Leonardo da Vinci. In 1593, Giacomo Porta (Italian architect, student of Michelangelo) established that in our minds are combined images obtained with both eyes, and described the individual components of a stereo pair.

The first devices allowing to obtain a stereoscopic effect appeared by the middle of the 19th century. In 1831, the first slit stereoscope was invented. It did not use optics, and the left and right images had to be rearranged. Later, in 1833, the first mirror stereoscope was manufactured, and in 1849 a lens stereoscope was developed. At this time, the basic principle of creating a stereo was already clear - the right eye sees the right frame of the stereo pair, and the left one sees the left frame.

Currently, stereo is gradually penetrating into many areas of human activity. An example of the most widely known application of stereo computer systems to the general public is stereo cinema.

Compared to conventional television, surround television has two new qualities. This is three-dimensional interactivity, which allows the viewer to be a co-creator of a television program, and three-dimensional image, which allows a person's eyes to work in a natural mode, moving their gaze from close objects of observation to distant ones. But this is only the beginning of the active introduction of stereo systems.

Recently, the number of professional stereo applications has grown steadily. For example, Vrey has developed an application program for translating CAD-created architectural projects into stereo images.

The additional sense of depth allows the most realistic to imagine of how the developed object will look in relation to the surrounding buildings.

Further development of the technology made it possible to combine functionality in one system that allows to combine the output of 3D and

conventional video. Thanks to the jewelery precision of the new technology, 3D projection video systems have begun to be used not only in entertainment, but also in precision engineering, automotive, aviation, geology, as well as in those areas where it is necessary to work with accurate three-dimensional models. Sitting behind the wheel of a car not yet embodied in metal, imagining and calculating the behavior of an aircraft or a rocket under aerodynamic action, plunging into the rock mass to lay the path of an oil well - all this became possible.

Also, one cannot fail to mention the most important area for humans, such as medicine. Many areas, for example, tomography, surgery, radiology, ophthalmology, ultrasound examination, need to be monitored and analyzed in a volumetric image. Programs with the use of stereo technologies have already been developed and tested, which are used to identify pathologies of the organs of vision and their treatment.

- a package of diagnostic programs "Diagnose", containing fifteen tests for measuring important parameters, designed to control binocular and stereoscopic vision;
- program "SHOW", which uses stereo animation as a simple test for the presence of stereoscopic perception by patients.

But despite the rapid development and use of stereo technologies, methodological and psychophysiological aspects of creating a virtual space and human interaction with it are still poorly developed, and there are many questions here.

Thus, the creation of an experimental setup, modeling and study of the processes of stereovision will make it possible to more fully study the peculiarities of the functioning of the human organ of vision in stereoscopic perception of the presented images.

The obtained results can be used to determine the requirements for video electronics systems operating in stereo mode and to optimize their geometric and optical parameters.

CHAPTER 2. ANALYSIS OF EXISTING ALGORITHMS FOR SOLVING THE PROBLEM

At the moment, there are many methods for finding out the range or distance to a body, but each of them has certain pros and cons. Thus, it is worth considering each of them and making sure of this.

2.1. Calculating the distance to an object from a photograph

As it is known, the simplest lens for a camera can be made from one biconvex lens. Of course, there are cameras without a lens at all (the so-called pinhole cameras, the ancestor of which is a pinhole camera), but in this case it doesn`t have any interest for developer. To begin with, it will be considered how an image is built in a simple single-lens lens [6], and then it will be shown that the same methods are well suited for complex lenses that combine more than a dozen consecutive lenses.

Let remind the diagram of the path of rays in a thin lens (Fig.2.1) from the school course in geometric optics:

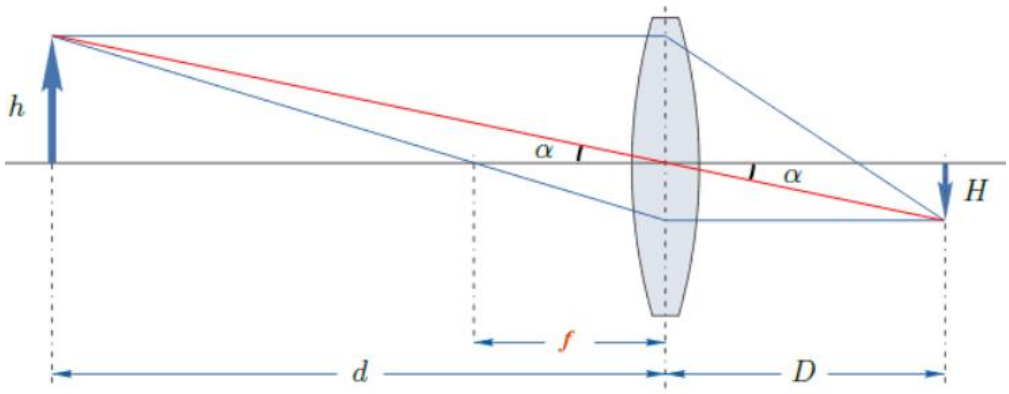

Fig. 2.1. Beam path in a thin lens

In this diagram, d is the distance from the lens to the object, D is the distance from the lens to the image of the object (on a matrix or film), and f is the focal length of the lens.

The thin lens formula from the same course consists these three distances:

$$\frac{1}{d}+\frac{1}{D}=\frac{1}{f}. \quad (2.1)$$

Now let's take another look at the optical scheme: h is the linear size of the subject, and H is the size of its reduced image. It is easy to see that $h = d * tan(\alpha)$, and $H = D * tan(\alpha)$ (this follows from the properties of a right-angled triangle). Substituting these values into the thin lens formula, it will be seen that $tan(\alpha)$ cancels out, and as a result, will obtain the following equation:

$$1+\frac{h}{H}=\frac{d}{f}. \quad (2.2)$$

The "inconvenient" value D is gone, and the rest are known or can be easily calculated. Based on this equation, get the following formula for the distance to the object:

$$d=\frac{f(H+h)}{H}. \quad (2.3)$$

After formulating the technical part, the proof can be seen in practice. Let's say there is a photograph of a house (Fig. 2.2), which is visible in full size. What useful information can be extracted from this photo? Let me remind that for the calculation it is needed unknown quantities h, H and f. h is the real height of the house (in meters). Immediately it is not known, but it can be found out this: the height of the ceilings in this house is 2.64 m, and the thickness of the floors is

0.22m. Surely, when measuring the height of the ceilings, the thickness of the floor covering was not taken into account. It is not known exactly, so, rounded a little, let’s take the height of one floor equal to 2.9 m. 23 panels are clearly visible, thus, the height of the visible area is about 66.7 m. Let's remember this value and start analyzing the photograph.

Fig. 2.2. Photo of the house

H is the size of the image of the house on the camera sensor. From a photograph, it can be calculated in pixels, but, as known, no one has exact data on the pixel size. But here it is necessary to remember that the camera matrix has specific physical dimensions. With the help of a search engine, find out that for the Nikon D90 camera, the matrix size is 2.36 × 1.58 cm, and the resolution is 4288 × 2848 pixels. The photo was not cropped or rotated, so it can be found out the exact linear size of the image of the house on the matrix by composing the proportion. But in order to do this not by hand, most often it is used Adobe Photoshop, which has a lot of useful tools.

Knowing the physical size of the matrix and the number of pixels along the long side of the image, make the following calculation: 4288 / 2.36 (matrix size in

centimeters), and get the correct resolution - 1817 pixels / cm. Enter it in the appropriate window and so that the actual dimensions of the photo do not change, but only its length and width in cm were recalculated. At the same time, the already known dimensions of the matrix appeared in the "Print size" field: 2.36 × 1.57 cm. More precisely, the specification indicated 1.58 cm, but this is an insignificant error. Now, using the Ruler tool, measure the height of the visible area of the house (23 panels, Fig. 2.3) in the photo.

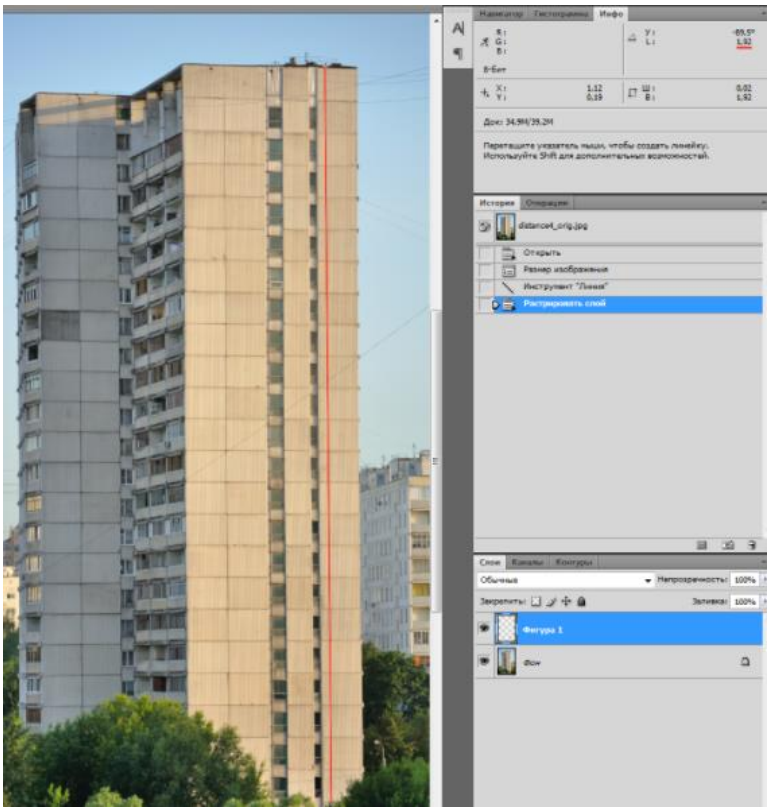

Fig. 2.3. Height of the visible area of the house

It turns out that the height of the image of the house on the matrix is 1.92 cm, or 0.0192 m. It remains only to find out the focal length, but for this, fortunately, it can be counted nothing: it is immediately registered when shooting in the metadata of the photograph (EXIF). Then, open them in a photo editor and see that the focal length during shooting was 105 mm, or 0.105 m, that is, the house was filmed with the maximum possible approximation for this lens. Now, all the data are ready for the calculation. Substitute them in formula 3 and get:

$$d = \frac{\left(0{,}105(0{,}0192 + 66{,}7)\right)}{0{,}0192} = 364{,}9\ \text{m}$$

The disadvantages of this method are the cost of data processing, since it is necessary to know in advance the certain dimensions of the photo area, which is not very correct in terms of the multitasking of the program, the more important is the fact that this method is based on any photographs of relatively large objects taken from a distance. more than 10 m, the result will be extremely accurate, but if the bodies are closer than this distance, an accuracy error occurs, which is not acceptable for work.

2.2. NanoLoc technology

An important question in determining the distance to the object: "Who determines the distance?" In the GPS system, this is done by a local device that positions itself on a grid, while, for example, RFID technology allows to determine the location of the object itself from the side; at the same time, the object itself does not have the ability to localize itself in space.

Convenient for the development of location technology is the ability to use the infrastructure of local wireless and wired networks. This has contributed to the introduction of a number of commercial solutions to the market based on such widespread standards as Bluetooth and Wi-Fi.

NanoLOC [7], also developed by Nanotron Technologies GmbH, automatically determines the distance between transmitting and receiving nodes. The developers declare the accuracy of measuring distances up to 1 meter. The NanoLOC technology uses one of the modifications of the above-mentioned Time of Flight method, called Symmetric Double Sided Two Way Ranging. SDS-TWR is a further enhancement of the Round Trip Time (RTT) method.

To measure distances by the RTT method between objects A and B, object A sends to object B a packet containing a measurement request and records the time of sending. Object B, having received a packet from A, sends an

acknowledgment to A - an ACK packet. Object A, having received an ACK packet, records the time of its receipt. RTT uses hardware ACK packet generation, where packet processing times are assumed to be the same for both objects. Fixing the time of sending a packet containing a request for measuring and receiving an ACK packet is also done in hardware. This allows to predict the packet processing time in advance and calculate the signal propagation time t_p using the formula:

$$t_p = \frac{T_{RTT} - T_{reply}}{2}. \quad (2.4)$$

where T_{RTT} - the time measured by object A from the moment the packet was sent to object B until the receipt of the ACK packet from object B; T_{reply} - time measured by object B from the moment of receiving the packet from object A to sending the ACK packet. Considering the speed of propagation of a signal in a medium as a known and constant value, it is easy to calculate the distance between objects. The accuracy of measuring time intervals, and, consequently, distances, when using NanoLOC technology [8], is significantly affected by the frequency stability of crystals used in transceiver modules. The degree of accuracy is characterized by the error ppm (part per million), for convenience written in integers. A value of 1 ppm corresponds to an error of 0.0001%. For example, for a crystal with a nominal frequency of 4 MHz and a frequency stability of 1 ppm, the maximum frequency deviation from the nominal will be 4,000,000 × 106 = 4 Hz.

Based on the experiments, the developers made the main conclusion: the used radio modules of the NanoLOC standard allow to accurately measure distances. At the same time, the measurement results are slightly overestimated (in different ways depending on external conditions). However, in some cases this can be corrected by preliminary calibration of the system or by introducing special calibration curves. Most often, the "raid of distances" is associated with reflections from neighboring objects, the existence of obstacles in the path between the sensors

and the different relative positions of the antennas. The minimum overestimation was achieved in experiments with a long time of data accumulation from stationary objects and amounted to 0.11 m.

2.3. Infrared sensor

An infrared motion sensor [9] is an electronic device capable of responding to changes in the intensity of background heat radiation in its area of action. Absolutely any object has thermal radiation, not just a person. Any object whose temperature is not lower than the air temperature emits heat. The task of the infrared sensor is to highlight it against the general thermal background and, when moving in a given area, send a signal.

A person is a thermal object rather "hot" compared to the walls of an apartment or house, the ground or trees. This allows the sensor to distinguish it from the general background.

When a heat object of suitable size and temperature crosses the serviced sector, the meter registers movement. The sensor then sends a signal to the control unit. Depending on what the working device with the sensor is intended for, the control module turns on the light, activates the security alarm, and so on.

The operating area of the registrar is limited. The detection radius should reach all corners of the room or to the end of the zone in the garden. If this is not the case, mount 2 or more meters.

The Kinect technology [10] works the same way (Fig. 2.4). Kinect (formerly Project Natal) is a touchless game controller originally introduced for the Xbox 360 console, and much later for the Xbox One and Windows PCs, developed by Microsoft. Based on the addition of a peripheral to the Xbox 360 game console, Kinect allows the user to interact with it without the aid of a contact game controller through verbal commands, body postures, and displayed objects or drawings.

Fig. 2.4. Kinect technology

The Kinect is a horizontally positioned box on a small circular base that is placed above or below the screen to interact with the console. The depth sensor consists of an infrared projector combined with a monochrome CMOS sensor, which allows the Kinect sensor to acquire a three-dimensional image in any natural light.

The depth range and project program automatically calibrates the transducer based on playing and environmental conditions such as furniture in a room.

The main disadvantage of this development is direct interaction only with a certain kind of software and devices for a number of applications, which entails the problem of high cost and difficulties for creating own program based for certain functions.

2.4. LIDAR

LiDAR is a remote sensing technology [11] that uses a laser pulse to collect measurements, which can then be used to create 3D models, maps of objects and the environment.

The technology has been around since the 1960s, when laser scanners were installed on airplanes. It wasn't until the late 1980s, with the advent of commercially viable GPS systems, that LiDAR data became a useful tool for

providing accurate geospatial measurements. LiDAR stands for Light Detection and Ranging.

It works similarly to radar and sonar, but uses light waves from a laser instead of radio or sound waves. The LiDAR system calculates how long it takes for light to hit an object and reflect back to the scanner. Distance is calculated using the speed of light.

The speed of light is 299,792,458 meters per second. The systems can generate about 1,000,000 pulses per second. Each of these measurements or results can then be converted into a 3D visualization, which is a point cloud.

Using Lidar technology. The system is most often used for surveying tasks. Due to their ability to collect 3D measurements, laser scanning systems have become actively used for surveying built environments (for example: buildings, road networks and railways), as well as for creating digital terrain models (DTM) and digital elevation model (DEM).

Laser scanning is a popular method for detecting flood risk, forestry carbon build-up, and monitoring coastal erosion. There is also an increased adoption of automation applications using this technology. Many car manufacturers use shorter range and lower range scanners to aid in the navigation of autonomous vehicles. With the use of this technology the automatic control systems which consists in Tesla cars and other types of such models work.

Today, the most common uses for the lidar system are in geographic and atmospheric mapping applications. Organizations such as the USGS (United States Geological Survey), NOAA (National Oceanographic and Atmospheric Administration) and NASA have used lidar for decades to create maps of the Earth and space.

Climatologists use it to study the composition of the atmosphere and study clouds, evaporation and global warming.

Oceanographers use it to track coastal erosion.

Botanists use LiDARs to measure the ever-changing structures of the Earth's forests.

Lidar can be also used to study the gas composition of the atmosphere. Different gases absorb light waves of different wavelengths in the required amount, so it can be remotely studied gases in a specific place by launching two laser beams with different wavelengths from an airplane or helicopter into it, then comparing how much of each wavelength is absorbed or reflected. This system, called Differential Absorption LiDAR (DIAL), can be used for everything from detecting leaks in gas lines to measuring air pollution.

One of the most common uses is police equipment for measuring the speed of vehicles, although people usually think that it is radar. Portable devices are much more likely to use 905 nm lasers, which are cheap, safe and very efficient.

LiDARs have a great future, as this technology does not stand still, constantly developing applications and utilities. From basic sensor applications to 3D printing systems, 3D scanning, simulation and smart cities. Lidar transforms the world in many ways.

Lidar in augmented reality. LiDAR Augmented reality is a technology that allows the user to view virtual content in the same way as it would exist in the real world. LiDAR enhances the clarity and end result of AR systems. The LiDAR scanner offers high quality "3D mapping" that allows other AR systems to overlay data on top of a high resolution map using a point cloud to complement it well.

Research is also underway on the application of "Doppler wind", which would make it possible to clearly see the movement of the wind. This approach would be very useful for aviation security, atmospheric data visualization, weather forecasting, and disaster preparedness.

LiDAR Mapping. The mapping uses a laser scanning system with a built-in Inertial Measurement Unit (IMU) and a GNSS receiver that enables georeferencing of each measurement or point. Each point is combined with others to create a three-dimensional representation of an object or area (Fig. 2.5).

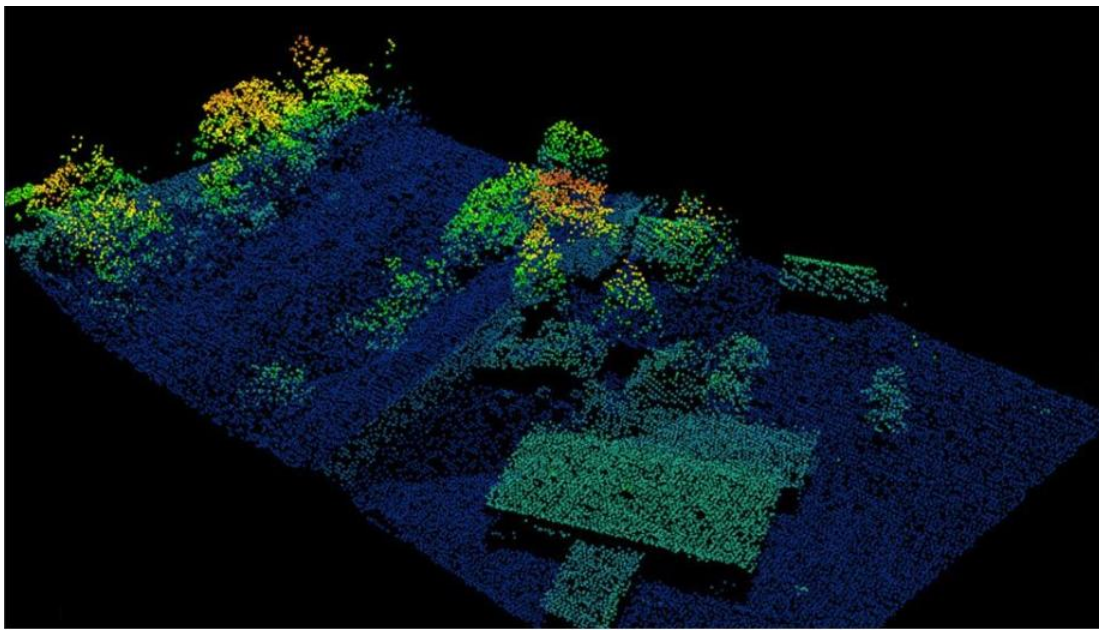

Fig. 2.5. 3D map created with LIDAR

Lidar maps can be used to determine positioning accuracy. LiDAR materials in the form or point cloud can be used to create maps of entire cities, with millimeter precision. Elements and objects such as road networks, bridges, vegetation can be classified and plotted on 3D maps. LiDAR maps can also be used to highlight changes and variances such as land erosion, changes in soil slope, and vegetation growth.

Lidar advantages:

- High speed and accuracy of collection;
- High penetration;
- Does not depend on the intensity of light in the environment and can be used at night or in the sunny weather;
- High image resolution compared to other methods;
- Lack of geometric distortion;
- Easily integrates with other collection methods;
- LIDAR has minimal dependence on humans, which is good in certain areas where human error is expensive.

Disadvantages of LiDAR:

- The cost of LIDAR is quite high;
- LIDAR systems do not work well in conditions of heavy rain, fog or snow;
- LIDAR systems generate large sets of materials that require large computing resources for processing;
- Unreliable results in turbulent water conditions;
- Depending on the wavelength adopted, the performance of the systems is limited in height, since the pulses generated in certain types become ineffective at certain altitudes.

2.5. Ultrasonic sensors

The ultrasonic rangefinder [12] determines the distance to objects in the same way as dolphins or bats do. It generates sound pulses at 40 kHz and listens for echoes. By the time of propagation of the sound wave back and forth, it is possible to unambiguously determine the distance to the object.

Unlike infrared rangefinders, ultrasonic rangefinder readings are not affected by light from the sun or the color of an object. However, it can be difficult to determine the distance to fluffy or very thin objects. Therefore, it will be difficult to perform a high-tech task on it.

An echo is heard when sound bounces off an obstacle. The bat uses the reflection of ultrasonic waves to fly in the dark and to hunt insects. An echo sounder works according to the same principle, with the help of which the depth of water under the bottom of the ship is measured or the search for fish.

The principle of transmitting and receiving ultrasonic energy is at the heart of many very popular ultrasonic sensors and speed detectors. Ultrasonic waves are mechanical acoustic waves, the frequency of which lies beyond the audibility of the human ear - more than 20 kHz.

The accuracy of the sensor depends on several factors:

- air temperature and humidity;
- distance to the object;
- location relative to the sensor (according to the radiation diagram);
- the quality of performance of the elements of the sensor module.

The principle of operation of any ultrasonic sensor is based on the phenomenon of reflection of acoustic waves propagating in the air. But as it is known from the course of physics, the speed of sound propagation in air depends on the properties of this air (primarily on temperature). The sensor, emitting waves and measuring the time until their return, does not guess in which medium waves will propagate and takes a certain average value for calculations. In real conditions, due to the air temperature factor, HC-SR04 may have errors from 1 to 3-5 cm.

The distance factor is important because the probability of reflection from neighboring objects increases, besides, the signal itself attenuates with distance.

To reduce errors and measurement errors, the following actions are usually performed:

- the values are averaged (measure it several times, remove the bursts, then find the average);
- using sensors (for example, DHT11 or DHT22), the temperature is determined and correction factors are introduced;
- the sensor is mounted on a servo motor, with which we “turn our head” by moving the directional pattern to the left or right.

Also, to improve accuracy, need to correctly direct the sensor: make sure that the object is within the cone of the directional pattern. Simply put, the HC-SR04 must be looking directly at the subject.

This method is not very suitable for most production scenarios due to the factors described earlier.

2.6. Technology for constructing 3D models of objects from a set of images

Today there is a whole set of software products for building 3D models of objects and scenes from sets of images [13] (for example, 123D Autodesk or Photomodeller). Now the general methodology for solving this issue will be described, with the capabilities of each of the stages.

First, it will be described the requirements for photographing an object (see Fig. 2.6). The overlap between a pair of frames of the photographed area of space should be no worse than 50% (otherwise the model will be broken). Moreover, such a survey should ensure that three adjacent images are overlapped (for example, in Figure 2.6, adjacent images can be considered 1,2,3 or 4,5,6). Thus, the resulting 3D model will be determined by only one scale parameter.

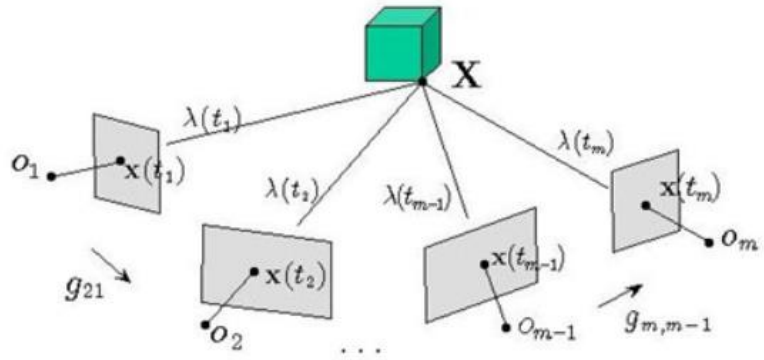

Fig. 2.6. The scheme of photographing the object

Now, a set of snapshots is done (see Fig. 2.7). Further, performing image processing (namely, searching for identical points of the object in the images and solving a system of nonlinear equations based on the found matches), determine the camera parameters (focal length, etc.) and the position / orientation of the camera at the moments of photographing each of the pictures relative to one of them.

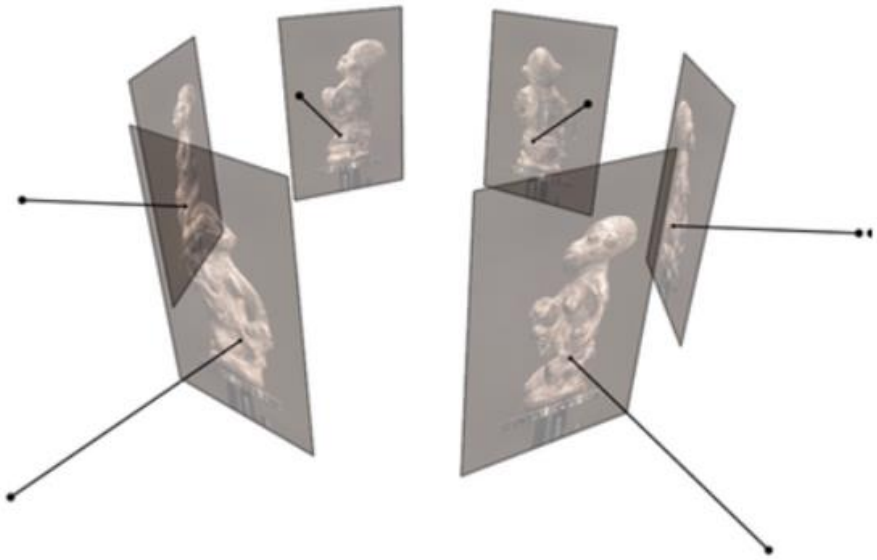

Fig. 2.7. The orientation of the snapshot set

For oriented images, all identical points are searched for on adjacent pairs of images (so-called dense maps or depth maps), after which the position of points in space (see Fig. 2.8) is calculated in the coordinate system of the base image (based on the calculated camera parameters: focal length, position / orientation, etc.).

Fig. 2.8. 3D model of the object

As a rule, many points are represented in the form of triangulation wireframes (the wireframe is built on the basis of Delaunay triangulation),

convenient for subsequent texturing (for example, using OpenGL tools) or transforming images.

Although this method is of high quality, but in itself is quite costly, in terms of the fact that in order to have a 3D scene, a lot of pictures always needed, which is not very fast and convenient for the construction of a stable structure of the device, since image capture will not be carried out only for a specific subject, then each time the cameras will need to be moved to obtain data, which can still be used later.

2.7. The method with the use neural networks

In recent years, there has been a rapid development of models of artificial neural networks, in particular, convolutional neural networks [14, 15], in which the processes of preprocessing and feature extraction are a consequence of their mechanism of operation. However, the operation of such networks requires the presence of pairs "training example - desired network output", which in the case of stereo images leads to the need for preliminary selection of depth maps. This requires an additional process of calibrating the camera in order to determine the external and internal parameters of the camera. Accordingly, if exclude this process from the technological chain when restoring a three-dimensional scene, then a significant simplification of calculations is possible. This approach is proposed to be implemented using the unsupervised learning method, which requires only the availability of training examples.

One of the unsupervised learning methods is the autoencoder, which allows to efficiently find dependencies in the input data. In this paper, we consider the models of the so-called convolutional autoencoders, which are more suitable for image processing tasks.

An important property of this class of neural networks is that in the process of training, compressed representations of input data are formed in hidden layers, which allows, for example, cleaning images from noise or generating new images. The works show the use of autoencoders when reconstructing depth maps from

single images; variants of the loss function for assessing the quality of reconstruction are given. In a study was carried out of the influence of the topology parameters of a multilayer perceptron on its operation, key parameters were identified that affect the convergence of perceptron learning, the proposed approach can be applied to convolutional neural networks. In the study of convolutional neural networks, it was found that increasing the depth of the network, with a relatively small number of convolution filters used, and together with the use of subsampling layers, leads to an improvement in the quality of classified images with a reduced computational load.

From research with neural networks, it can be seen that:

- autoencoders restore the input image by compressing transformation, due to which the essential features are extracted;
- an increase in the number of filters in convolution layers does not lead to a significant increase in accuracy, but leads to a significant increase in the number of calculations, the opposite is also true;
- at low dimensions of feature maps at the border of the encoding and decoding parts, the solution accuracy is significantly deteriorated, in which, in addition to reducing the quality of the reconstructed image, color information is also lost;
- autoencoder shows effective results even with shallow network depth, which increases the speed of image processing. This is confirmed by the results of different types of model, which, with a relatively small number of training parameters, demonstrated the best results after 90 iterations of training on test data.

2.7.1. Constructing with Capsule Neural Networks

Convolutional neural networks are capable of registering only the presence of an object in a picture [16], without encoding its orientation and position. But capsular neural networks do not have this disadvantage.

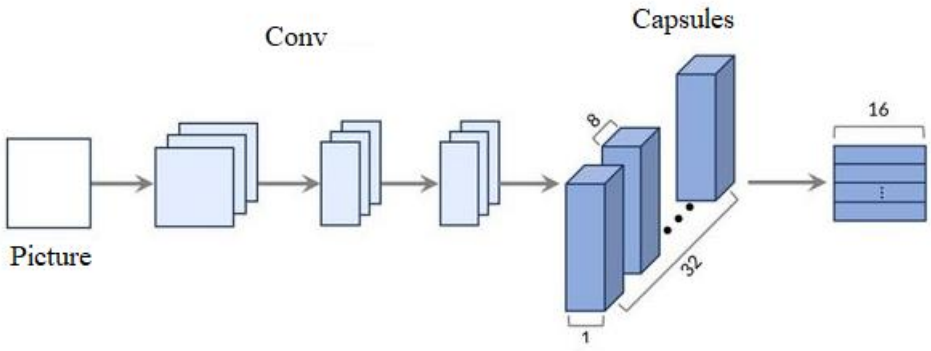

Fig. 2.9. Capsular neural network structure

A CapsNET consists of capsules or groups of neurons to identify patterns in an image. This information comes in the form of vectors containing the orientation and position of the patterns in the image, which is then received by the higher-level capsules. The higher-level capsules process this information from several lower-level capsules and then generate a prediction. Capsules of the same level have no connections with each other and calculate information independently of each other. Capsules are formed by splitting the output from the convolutional layer. Divide the three-dimensional vector into capsules using the "slicing" method so that each capsule contains information about each pixel, i.e. in three-dimensional coordinate.

The state of the neurons of the capsular neural network inside the image fixes the property of an area or object within the image: its position and orientation.

Using a capsule neural network is similar to using conventional convolutional networks described above. In general, this network shows more accurate depth prediction results.

2.7.2. Constructing with PlanetNet

There are also architectures that solve this problem without training on a disparity map built using two images. One of these is PlaneNet.

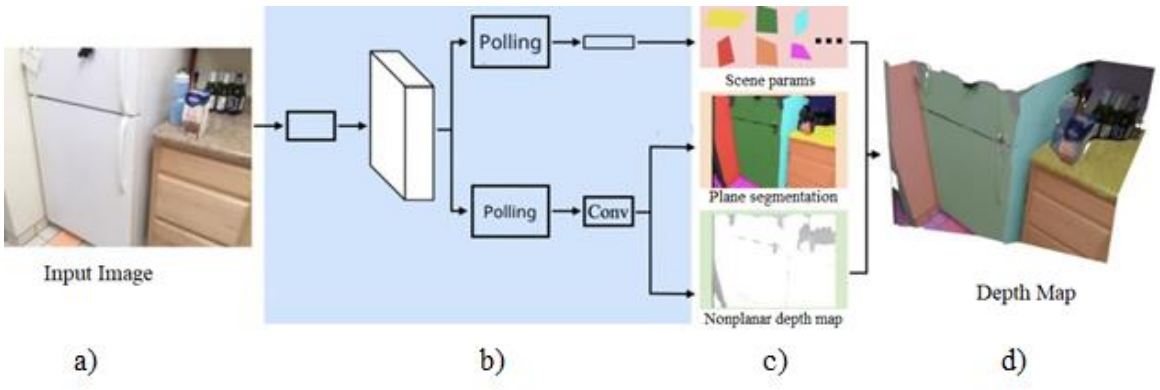

Fig 2.10. Predicted PlaneNet parameters from one RGB picture: a) Input image; b) PlaneNet layers; c) Plane segmentation, plane parameters, non-planar depth map; d) Output depth map image

PlaneNet is a deep neural network built on Dilated Residual Networks (DRN). It obtains a depth map by composing the outputs of three subtasks (Fig. 2.10):

- Parameters of planes: trying to predict the number of planes K, and then looking for K flat surfaces on the image, each surface is set by three parameters P_i: normal, straight and shear. The error function is defined as follows:

$$L = \sum_{i=1}^{K} \min_{j \in |1,\hat{K}|} ||\hat{P}_j - P_i|| , \tag{2.5}$$

where $\hat{K}, \hat{P}_i,$ and $K, P_i,$ predicted and real number and parameters of planes, respectively.

- Segmentation of the plane: looking for groups of pixels, each of which characterizes one semantic object. Use cross entropy as a loss function.
- Non-flat depth map: looking for a single-channel (or non-flat) depth map, that is, a depth map where each pixel is either at depth 0 or at depth 1.

Most neural networks(NN) show a high-quality result, but there are still drawbacks, since NN require a large number of training samples. Also, the training time for such models requires a fairly long period. In addition, in order to use and

train a neural network, the PC system requirements must be high enough, and this is not always possible and requires a lot of money.

2.8. Radar method

Radar is a field of radio electronics dealing with the detection of objects (targets), the determination of their spatial coordinates, movement parameters and physical dimensions using radio equipment and methods.

The listed tasks are solved in the process of radar surveillance [17], and the devices designed for this is called radar stations (RS).

Radar targets (or simply targets) include: manned and unmanned aerial vehicles, natural and artificial space bodies, atmospheric formations, sea and river ships, various ground and underground, surface and underwater objects, etc.

Target information is contained in radar signals. In the case of radar sounding of aircraft, first of all, it is necessary to obtain information about their spatial coordinates (range to the target and its angular coordinates).

Radio-technical measurements of range are called radio range measurements, and angular coordinates are called radio direction-finding.

The measurement of the coordinates and speed of targets is preceded by their detection, resolution and identification. Goal resolution is understood as determining the number of goals in a group, their length, class, etc. Identifying a goal means establishing its essential features, in particular, nationality. The definition of the type (class) of the target is carried out in the process of its recognition, which implies complex processing of radar signals.

The collection of information received by radar facilities is called radar information. The latter is transferred to command posts, calculating devices and executive devices.

Of all the listed radar functions, the main one is radar surveillance (target detection, measurement of coordinates and motion parameters), and object discrimination, identification of them and transmission of the received radar information to the destination are additional radar functions.

Receiving radar information is based on the physical properties of electromagnetic waves (EMW) used as carriers of the radar signal. As it is known, EMWs propagate in a homogeneous medium in a straight line with a constant speed:

$$v = \frac{1}{\sqrt{\varepsilon_a \mu_a}}, \tag{2.6}$$

where ε_a, μ_a - absolute dielectric and magnetic permeability of the targets medium (for free space $\varepsilon_a = \varepsilon_0 = \frac{1}{36}\pi \cdot 10^9$ F/m; $\mu_a = \mu_0 = 4\pi \cdot 10^{-7}$ H/m and correspondingly $v = c = 3 \cdot 10^8$ m/sec.

The constancy of the propagation velocity vector of EMW in a homogeneous medium, i.e., its modulus and direction, serves as a physical basis for radar measurements.

Indeed, due to this, the range D and the propagation time of the radio wave (RW) are related by direct proportionality, and if the wave propagation time is measured t_d between the target and the radar, then the distance between them becomes known:

$$D = c * t_d. \tag{2.7}$$

The goal introduces heterogeneity into the free space, since its parameters ε_a and μ_a differ from ε_0 and μ_0, respectively, which violates the constancy of the velocity vector of the radar. As a result, the object converts radio emission: part of the energy is re-reflected, part is absorbed by the object, turning into heat, and the other part, when the object is radio-transparent, is refracted, changing the direction of the RWR. From the point of view of radar, the first case is interesting when the target becomes a source of secondary radiation. By the time delay of the reflected signal relative to the emitted one:

$$t_d = \frac{2D}{c} \tag{2.8}$$

determine the slant range of the target

$$D = \frac{ct_d}{2}. \tag{2.9}$$

Such a solution is also possible: a transceiver called a transponder, or a repeater, is installed on the target, if it is "one of our own", and not the enemy, which receives the sounding signal from the radar and amplifies it to start the transmitter. The response signal is received by the radar, and the target range is determined by the formula

$$D = \frac{c(t_d - t_{ans})}{2}, \tag{2.10}$$

where t_d - delay of the response signal relative to the probe; t_{ans} - known in advance the delay time of the signal in the transponder circuits.

The quantity t_d should be measured by an inertialess electronic clock, since the delay time of radar signals is very small (from micro to milliseconds).

For example, radar reflected from a target located at a distance of D = 150 m from the radar are delayed by 1 μs, and each kilometer of range to the target corresponds to a radar delay for a time of 1000/150 = 6.7 μs.

The radial and angular velocities of the target can be found by calculating the rate of increase in range and angles over time. Usually, a simpler and more accurate operation is preferred - direct measurement of the so-called Doppler shift of the carrier frequency of the signal f_0, caused by the movement of the target.

Doppler frequency offset F_D associated with the radial speed of the object V_r ratio

$$F_D = -\left(\frac{2V_r}{c}\right) f_0 = -\frac{2V_r}{\lambda_0}, \quad (2.11)$$

where λ_0 - the wavelength of the emitted signal; V_r - the radial velocity of the relative movement of the target.

If the target approaches the radar or moves away from it, then the reflected signal appears in the radar, respectively, earlier or later than when the target is stationary. Due to this, the phase of the received wave has different values, which is equivalent to the increment in the frequency of the radio signal. By measuring the obtained (Doppler) frequency increment, it is possible (again due to the constancy of the radar speed) to determine the radial velocity of the target.

Just as the difference in the signal lag time in the antenna elements is determined by the angular coordinates of the target, the difference in the Doppler frequency offset in the same (usually extreme) antenna array elements is determined by the rate of change in the angular position of the target.

Other physical properties of EMW are:

- straightness of propagation in a homogeneous medium, which is important for accurate measurement of angular coordinates and motion parameters;
- the ability to form into a narrow beam, thereby increasing the accuracy, resolution and noise immunity of the radar;
- ability to reflect from objects; the ability to change its frequency in the presence of relative movement of the target and radar.

Thus, the radar signals reflected from the targets contain all information about them, since all signal parameters (amplitude, frequency, initial phase, duration, spectrum, polarization, etc.) change during reflection. But, however, these advantages come at the cost of increased complexity and cost of the system.

It becomes necessary to synchronize the work of positions (including when viewing space) and to organize data transmission lines. The complexity of information processing also increases due to its large work content.

CHAPTER 3. THEORETICAL INFORMATION ABOUT THE SOLUTION OF THE PROBLEM

Due to the fact that when solving this problem, two stereo cameras algorithm was chosen [18], it is worth saying that it can be told about determining the three-dimensional coordinates of the observed points when there are at least two cameras. Namely, it represents the detection of principle points, as shown in Fig. 3.1.

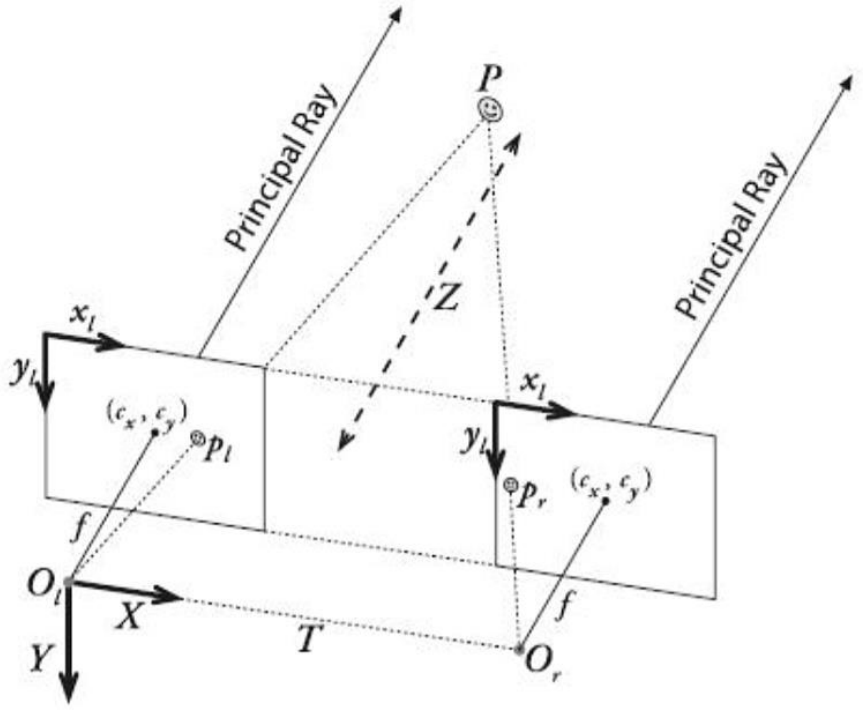

Fig. 3.1. Stereo coordinate system

In the image, it can be seen that two cameras are parallel to each other and at the same height with the centers of the optical lenses O_l, O_r, it is also important that the focal length of the planes coincide, which is done during the settings and correct installation of the cameras, P is the object that represents interest for the observer, Z is the real distance that can be measured in any suitable length measures.

But it should be borne in mind that to measure the distance to an object at a large distance, the distance between the optical centers of the lenses T must be increased, but most often, objects in within 15 meters, and for this, it can be used an analogue with human vision, where the average distance between the pupils is 55 millimeters.

Since the maximum analogue to human vision was used, it is worth explaining the essence of the work of the brain at this moment. The brain receives two different images from each eye, and perceives them as one three-dimensional image. Despite the fact that the image of objects on the retinas is two-dimensional, a person sees the world in three-dimensional, that is, can perceive the depths of space with stereoscopic vision. Man has many mechanisms for assessing depth. For example, if observer know the size of an object (man, tree, etc.), then person can determine the distance to it and find out which of the objects is closer by comparing the angular magnitude of the object. But if one object is located in front of the other and partially covers it, then the front object seems to be closer to the person.

The machine sees the same thing. Now it will be considered the principle of projective geometry on what the work is based on.

3.1. Projective geometry and homogeneous coordinates

In the geometry of stereo vision, projective geometry plays a significant role. There are several approaches to projective geometry: geometric (like Euclidean geometry, introduce the concept of geometric objects, axioms and from this deduce all the properties of projective space), analytical (consider everything in coordinates, as in the analytical approach to Euclidean geometry), algebraic. For the further presentation, an understanding of the analytic approach to projective geometry is mainly needed, and it is this approach that is presented below.

3.1.1. Points of the projective plane

Consider a two-dimensional projective space (which is also called the projective plane). Whereas on the ordinary Euclidean plane, points are described by a pair of coordinates $(x, y)^T$, on the projective plane, points are described by a three-component vector $(x, y, w)^T$. Moreover, for any nonzero number a, the vectors $(x, y, w)^T$ and $(ax, ay, aw)^T$ correspond to the same point. And the zero vector $(0,0,0)^T$ does not correspond to any point and is discarded from consideration. Such a description of points on the plane is called homogeneous coordinates.

The points of the projective plane can be associated with the points of the ordinary Euclidean plane. To the coordinate vector $(x, y, w)^T$ for w ≠ 0 associate a point in the Euclidean plane with coordinates $\left(\frac{x}{w}, \frac{y}{w}\right)^T$. If w = 0, i.e. the coordinate vector has the form $(x, y, 0)^T$, then this point is at infinity. Thus, the projective plane can be viewed as the Euclidean plane, supplemented by points from infinity.

It is possible to pass from homogeneous coordinates $(x, y, w)^T$ to ordinary Euclidean ones by dividing the coordinate vector by the last component and then discarding it $(x, y, w)^T \rightarrow \left(\frac{x}{w}, \frac{y}{w}\right)^T$. And from the Euclidean coordinates $(x, y)^T$ one can pass to homogeneous ones by complementing the coordinate vector with one: $(x, y)^T \rightarrow (x, y, 1)^T$.

3.1.2. Lines on the projective plane

Any straight line on the projective plane is described, like a point, by a three-component vector $l = (a, b, c)^T$. Again, the vector describing the line is defined up to a nonzero factor. In this case, the equation of the straight line will have the form: $l^T x = 0$.

In the case when $a^2 + b^2 \neq 0$ we have an analogue of the usual straight line $ax + by + c = 0$. And the vector $(0,0, w)$ corresponds to a straight line lying at infinity.

3.1.3. Three-dimensional projective space

By analogy with the projective plane, points of the three-dimensional projective space are determined by the four-component vector of homogeneous coordinates $(x, y, z, w)^T$. Again, for any nonzero number a, the coordinate vectors $(x, y, z, w)^T$ and $(ax, ay, az, aw)^T$ correspond to the same point.

As in the case of the projective plane, a correspondence can be established between the points of the three-dimensional Euclidean space and the three-dimensional projective space. The vector of homogeneous coordinates $(x, y, z, w)^T$ for w $\neq$ 0 corresponds to a point in Euclidean space with coordinates $\left(\frac{x}{w}, \frac{y}{w}, \frac{z}{w}\right)^T$. A point with a vector of homogeneous coordinates of the form $(x, y, z, 0)^T$ is said to lie at infinity.

3.1.4. Projective transformation

One more thing that will be required for further presentation is projective transformations. From a geometric point of view, a projective transformation is an invertible transformation of a projective plane (or space), which transforms straight lines into straight lines. In coordinates, the projective transformation is expressed as a nondegenerate square matrix H, while the coordinate vector x is transformed into the coordinate vector x 'according to the following formula: $x' = H\,x$.

3.2. Projection camera model

Modern cameras are well described using the following model called the projection camera. The projective camera is determined by the center of the camera, the main axis - the ray starting in the center of the camera and directed to where the camera is looking, the image plane - the plane onto which the points are projected, and the coordinate system on this plane. In such a model, an arbitrary point in X space is projected onto the image plane at a point x lying on the segment CX, which connects the center of the camera C with the original point X (see Fig. 3.2).

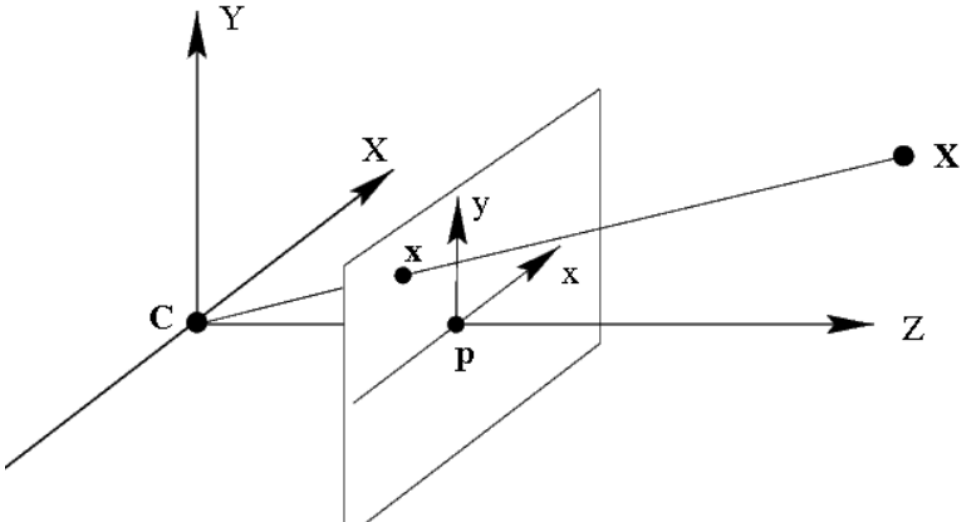

Fig. 3.2. Camera model. C is the center of the camera; Z is the main axis of the camera; The X point in 3D space is projected to the x point on the image plane

The projection formula has a simple mathematical representation in homogeneous coordinates:

$$x = P * X, \tag{3.1}$$

where X are homogeneous coordinates of a point in space, x are homogeneous coordinates of a point on a plane, P is a 3 × 4 camera matrix.

The matrix P is expressed as follows

$$P = KR[\,I\,|{-c}] = K[R|t], \tag{3.2}$$

where K is the upper triangular matrix of the internal parameters of the 3 × 3 camera (a specific view is given below), R is the 3 × 3 orthogonal matrix that determines the rotation of the camera relative to the global coordinate system, I is the 3 × 3 identity matrix, the vector c is the coordinates the center of the camera, and $t = -Rc$.

It is worth noting that the camera matrix is determined up to a constant non-zero factor, which will not change the results of projection of points by the formula

$x = P * X$. However, this constant factor is usually chosen so that the camera matrix looks as described above.

In the simplest case, when the center of the camera lies at the origin, the main axis of the camera c is directed to the Cz axis, the coordinate axes on the camera plane have the same scale (which is equivalent to square pixels), and the center of the image has zero coordinates, the camera matrix will be equal to $P = K[I|0]$, where

$$K = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.2)$$

In real video cameras, pixels are usually slightly different from square ones, and the center of the image has non-zero coordinates. In this case, the matrix of internal parameters takes the form

$$K = \begin{pmatrix} a_x & 0 & x_0 \\ 0 & a_y & y_0 \\ 0 & 0 & 1 \end{pmatrix}. \quad (3.3)$$

The coefficient f, α_x, α_y are called the focal lengths of the camera (respectively, general and along the x and y axes).

In addition, due to the imperfection of the optics, there are distortions in the images obtained from cameras. These distortions have a non-linear mathematical notation.

$$\begin{cases} x'' = x'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + 2p_1 x'y' + p_2(r^2 + 2x'^2) \\ y'' = y'(1 + k_1 r^2 + k_2 r^4 + k_3 r^6) + p_1(r^2 + 2y'^2) + 2p_2 x'y' \end{cases}, \quad (3.4)$$

where k_1, k_2, p_1, p_2, k_3 - distortion coefficients, which are the parameters of the optical system; $r^2 = x'^2 + y'^2$; (x', y') - coordinates of the projection of a point relative to the center of the image with square pixels and no distortion; (x'', y'') - distorted coordinates of a point relative to the center of the image with square pixels.

Distortions do not depend on the distance to the object, but depend only on the coordinates of the points to which the object's pixels are projected. Accordingly, to compensate for distortions, the original image obtained from the camera is usually converted. This transformation will be the same for all images obtained from the camera, provided the focal length is constant (mathematically, the same matrix of internal parameters).

In a situation when the internal parameters of the camera and the distortion coefficients are known, the camera is calibrated.

3.3. Interconnection of two cameras

When determining three-dimensional coordinates, matrices of a pair of cameras, namely their calibration, are of particular interest.

Let there are two cameras defined by their matrices P and P' in a certain coordinate system. In this case, it is said that there are a pair of calibrated cameras. If the centers of the cameras do not coincide, then this pair of cameras can be used to determine the three-dimensional coordinates of the observed points.

Often, the coordinate system is chosen so that the camera matrices have the form:

$$P = K[I|0], \tag{3.5}$$

$$P' = K'[R'|t']. \tag{3.6}$$

This can always be done by choosing the origin that coincides with the center of the first camera, and directing the Z axis along its optical axis.

Calibration of cameras is usually performed by multiple shooting of a certain calibration template; it is easy to select key points on the image, for which their relative positions in space are known. Further, systems of equations are compiled and solved (approximately) that connect the coordinates of projections, matrices of cameras and the position of the template points in space.

3.3.1. Epipolar geometry

Before proceeding to the description of the actual method for calculating the three-dimensional coordinates of points, we describe some important geometric properties that connect the positions of the projections of a point in three-dimensional space in the images from both cameras.

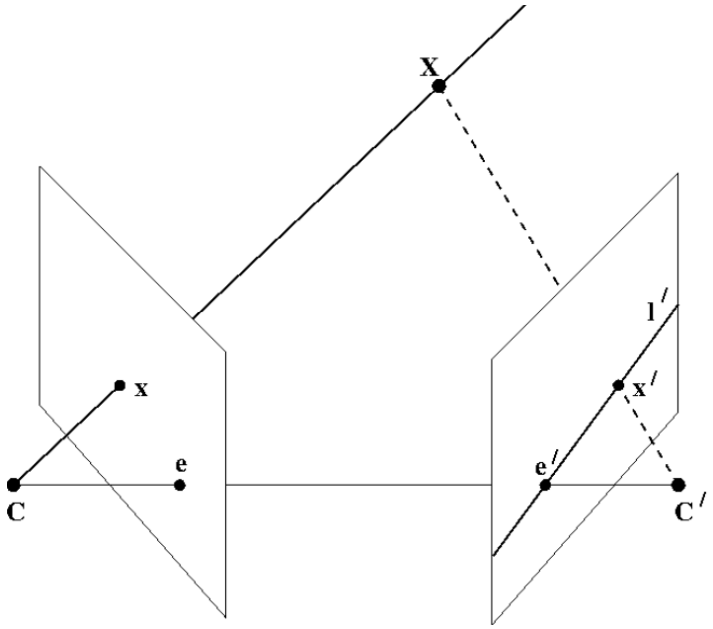

Fig. 3.3. Epipolar geometry

Let there are two chambers, as shown in Fig. 3.3. C is the center of the first chamber, C' is the center of the second chamber. The point in X space is projected at x on the left camera's image plane and at x 'on the right camera's image plane. Ray xX is the prototype of the point x in the image of the left camera. This ray is projected onto the plane of the second camera in a straight line l', called the epipolar line. The image of the point X on the image plane of the second camera necessarily lies on the epipolar line l'.

Thus, each point x in the image of the left camera corresponds to an epipolar line l' in the image of the right camera. In this case, the pair for x in the image of the right camera can only lie on the corresponding epipolar line. Similarly, each point x' on the right image corresponds to an epipolar line l on the left.

Epipolar geometry is used to search for stereopairs, and to check that a pair of points can be a stereopair (i.e. a projection of some point in space).

Epipolar geometry has a very simple coordinate notation. Let there be a pair of calibrated cameras, and let x be the uniform coordinates of a point on the image of one camera, and x' - on the image of the second one. There exists a 3 × 3 matrix F such that a pair of points x, x' is a stereopair if and only if

$$x'^{T} * F * x = 0. \tag{3.7}$$

The matrix F is called the fundamental matrix. Its rank is equal to 2, it is determined up to a nonzero factor and depends only on the matrices of the original cameras P and P'.

In the case when the camera matrices are of the form:

$$P = K[I|0], \tag{3.8}$$

$$P' = K'[R|t], \tag{3.9}$$

the fundamental matrix can be calculated by the formula

$$F = K'^{(-1)^{T}} R K^{T} [K R^{T} t] x, \tag{3.10}$$

where for the vector e (epipolar coordinates) the notation $[e]_X$ is calculated as

$$[e]_X = \begin{bmatrix} 0 & -e_z & e_y \\ e_z & 0 & -e_x \\ -e_y & e_x & 0 \end{bmatrix}. \quad (3.11)$$

Equations of epipolar lines are calculated using the fundamental matrix. For point x, the vector defining the epipolar line will have the form $l' = F * x$, and the equation of the epipolar line itself: $l'^T x' = 0$. Similarly for point x', the vector defining the epipolar line , will have the form $l = F^T x'$.

In addition to the fundamental matrix, there is also such a thing as an essential matrix:

$$E = K'^T * F * K. \quad (3.12)$$

In the case when the matrices of internal parameters are unit, the essential matrix will coincide with the fundamental one. Using the essential matrix, it is possible to restore the position and rotation of the second camera relative to the first, therefore it is used in tasks in which it is necessary to determine the camera movement.

3.3.2. Points triangulation

Now let's move on to how to determine the three-dimensional coordinates of a point by the coordinates of its projections. This process is called triangulation in the literature.

Let there are two calibrated cameras with matrices P_1 and P_2. x_1 and x_2 are the homogeneous coordinates of the projections of some point in the space X. Then it can be composed the following system of equations:

$$\begin{cases} x_1 = P_1 X \\ x_2 = P_2 X \end{cases}. \quad (3.13)$$

In practice, the following approach is used to solve this system. The vector is multiplied by the first equation by x1, the second by x2, get rid of linearly dependent equations and bring the system to the form $A * X = 0$, where A has a size of 4×4. Then it can be either proceed from the fact that the vector X is the homogeneous coordinates of the point, put it the last component is equal to 1 and solve the resulting system of three equations with three unknowns. An alternative way is to take any nonzero solution of the system $A * X = 0$, for example, calculated as a singular vector corresponding to the smallest singular number of the matrix A.

3.4. Depth Map constructing

A depth map [19] is an image in which for each pixel, instead of a color, its distance to the camera is stored. The depth map can be obtained using a special depth camera (for example, the Kinect sensor is a kind of such a camera), and it can also be built from a stereopair of images.

The idea behind building a depth map from a stereopair is very simple. For each point in one image, a search is performed for its paired point in another image. And for a pair of corresponding points, can triangulate and determine the coordinates of their prototype in three-dimensional space. Knowing the 3D coordinates of the preimage, the depth is calculated as the distance to the camera plane.

The paired point must be looked for on the epipolar line. Accordingly, to simplify the search, images are aligned so that all epipolar lines are parallel to the sides of the image (usually horizontal). Moreover, the images are aligned so that for a point with coordinates (x_0, y_0), the corresponding epipolar line is given by the equation $x = x_0$, then for each point the corresponding paired point must be searched for in the same line on the image from the second camera. This process of aligning images is called rectification. Rectification is usually performed by remapping the image and is combined with getting rid of distortions.

After the images are rectified, a search is performed for the corresponding pairs of points. The simplest way is illustrated in Fig. 3.4 and is as follows. For each pixel of the left picture with coordinates (x_0, y_0), a pixel is searched for in the right picture. It is assumed that the pixel on the right picture should have coordinates $(x_0 - d, y_0)$, where d is a value called disparity. The search for the corresponding pixel is performed by calculating the maximum of the response function, which can be, for example, the correlation of the neighborhoods of the pixels. The result is a disparity map.

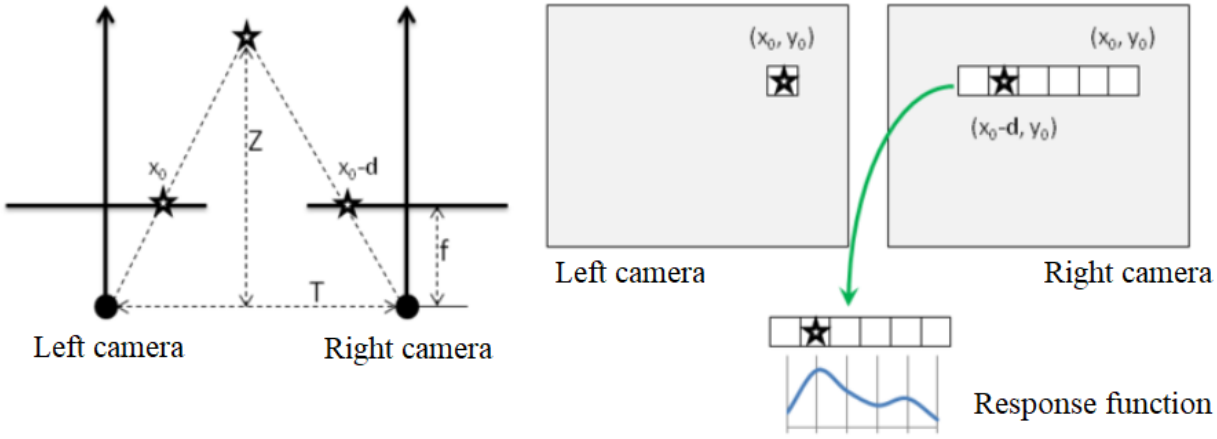

Fig. 3.4. Calculation of the depth map

Actually, the depth values are inversely proportional to the amount of pixel offset. Using the notation on the left half of Fig. 3.4, the relationship between disparity and depth can be expressed in the following way:

$$\frac{T-d}{Z-f} = \frac{T}{Z} \rightarrow Z = \frac{f * T}{d}. \qquad (3.14)$$

Due to the inverse relationship between depth and offset, the resolution of stereo vision systems that use this method is better at close distances, and worse at far.

3.5. Image filtering

In the process of transmission and conversion by means of radio engineering systems, images are exposed to various interference, which in some cases leads to

a deterioration in visual quality and loss of image areas. With the widespread introduction of digital communication systems, the actuality of solving the problems of recovering images obtained with the help of photo and video cameras, in order to filter images, increases. In practice, there are often images distorted by noise, which appears at the stages of its formation and transmission over the communication channel.

Computer graphics are divided into three main areas: visualization, image processing and pattern recognition. Visualization is the creation of an image based on a description (model). The main task of pattern recognition is to obtain a semantic description of the depicted objects. Image processing is responsible for transforming (filtering) images. The development of modern means of computer technology and information technology contributes to the widespread introduction of automatic image processing systems into practice.

The primary goal of such a system is to improve image quality. The problem of noise reduction is one of the most common problems in the field of image processing. The most common types of noise are Gaussian and impulse noise, and a combination of both.

The task of image processing can be either improvement of the image according to some specific criterion, or a special transformation that radically changes the image. In the other case, image processing can be an intermediate stage for further image recognition (for example, to highlight the contour of an object). Image processing methods can vary significantly depending on how the image was obtained - synthesized by a computer graphics system, or by digitizing a black-and-white or color photograph or video. In the case, if an image was obtained by digitizing them, usually, noise is present.

Most often, noise reduction serves to improve visual perception, but it can also be used for some specialized purposes - for example, in medicine to increase the clarity of an image on X-ray images, as a preprocessing for subsequent recognition, etc. Also, noise reduction plays an important role in image

compression. When compressed, a lot of noise can be mistaken for image detail, and this can adversely affect the resulting quality of the compressed image.

There are various sources of noise [20]:

1) imperfect equipment for image capture - video camera, scanner, etc.;

2) poor shooting conditions - for example, strong noises arising from night photo or video shooting;

3) interference during transmission over analog channels - interference from sources of electromagnetic fields, intrinsic noise of active components (amplifiers) of the transmission line.

Types of noise.

Accordingly, noises are also of different types. The models of additive Gaussian and impulse noise are the most adequate from the point of view of use in practical problems. Additive Gaussian noise is characterized by adding values from the corresponding normal distribution with zero mean to each pixel in the image. This noise is usually introduced during the digital imaging stage. Impulse noise is characterized by the replacement of some of the pixels in the image with values of a fixed or random value. Such a noise model is associated, for example, with errors in image transmission.

Noise removal methods.

Noise reduction algorithms usually specialize in suppressing a particular type of noise. There are still no universal filters that can detect and suppress all types of noise. However, many noises can be approximated fairly well by the white Gaussian noise model, so most algorithms are focused on suppressing this particular type of noise.

The most common methods for removing noise [21] are:

1. Smoothing filters;
2. Wiener filters;
3. Median filters;
4. Ranking filters.

Both linear and nonlinear filters are used to suppress Gaussian noise. A linear filter is defined by a real-valued function (filter kernel) defined on the raster. The filtering itself is performed using a discrete convolution (weighted summation) operation. With linear smoothing filtering, the intensity value at each point is averaged over a certain smoothing mask.

In the first case, the average value of the intensities of the neighbors is assigned to the intensity value at the central point. In other cases, a weighted average according to the coefficients.

Potentially better image processing results, in particular filtering results, are achieved when using a Wiener filter. Its application is associated with the assumption that the image is stationary. Since the presence of image edges serves as a violation of stationarity, the Wiener filtering is not strictly optimal. However, with frame sizes much larger than the image correlation interval, the effect of the edges is small. Technically, the Wiener filter is implemented using a discrete Fourier transform in the frequency domain.

But the use of linear filtering methods does not allow obtaining an acceptable solution in a number of practically important problems. It is necessary to take into account the nonlinear nature of the processes of transmission, coding and perception of information themselves, for example, information sensors, a communication channel, the human visual system, etc.

In order to expand the range of tasks solved by means of digital image processing, and to overcome the limitations inherent in linear filtering methods, nonlinear digital filtering methods [22] are currently being actively introduced.

Unlike the theory of linear filtration, the construction of a unified theory of nonlinear filtration is hardly possible. Each of the listed classes has its own advantages and scope. So, for example, it is known that the 9 best results for preserving the gradients of hues, various boundaries and local peaks of brightness in images distorted by impulse noise can be obtained by using median filtering.

The median filter, in contrast to the smoothing filter, implements a non-linear noise suppression procedure. The median filter is a window w sliding over the image field, covering an odd number of samples. The center count is replaced by the median of all image elements that fall into the window. The median of a discrete sequence is the average in the order of the term of the series obtained by ordering the original sequence.

Like the smoothing filter, the median filter is used to suppress additive and impulse noise in the image. A characteristic feature of the median filter, which distinguishes it from the anti-aliasing one, is the preservation of brightness differences (contours). Moreover, if the brightness differences are large compared to the variance of the additive white noise, then the median filter gives better results than the optimal linear filter. The median filter is especially effective in the case of impulsive noise.

The ranking filter [23], like the smoothing filter, uses a mask to transform the image. The mask may or may not include the center pixel. The values of the elements falling into the mask can be arranged in an ordered row and sorted in ascending (or descending) order, and certain moments of this series can be calculated, for example, the average value of intensity and variance. The output value of the filter that replaces the center sample is the weighted sum of the intensity of the center pixel and the median of the resulting series. The coefficients are usually related in some way to the pixel statistics in the filter window.

The algorithms described above allow to smooth out distorted images, gradually increasing their quality. Compared with linear and median filters, the combined and hybrid filters are the worst in terms of performance. This happens in connection with the calculation of the output of the subfilters (linear or median filters). In order to get rid of Gaussian noise in the image, the best solution is linear filtering, since such filters blur the details of the image itself. The easiest way to eliminate impulse noise is to use rank-based nonlinear filters while preserving all the jumps in the image. If there is a need to get rid of combined noise in the image,

then the easiest way to do this is using algorithms that take into account the peculiarities of this noise (combined, hybrid and adaptive filters).

CHAPTER 4. EXPERIMENTAL SOLUTION OF THE STATED PROBLEM

In order to solve the problem of recognizing the depth map, many step-by-step adjustments and adaptations for the research environment were done.

Each of these points will be discussed below.

4.1. Camera selection

Since, the interest was to make a cheaper and more effective version of the work, then, based on the qualities and selection at a price, two UTM Webcam (SJ-922-1080) web cameras were chosen (Fig. 4.1).

Fig. 4.1. Real View UTM Webcam (SJ-922-1080)

These cameras have such characteristics:

Video resolution - FullHD (1920x1080);

- Sensor - CMOS;
- Built-in microphone - With microphone;
- Focusing - Autofocus;
- Mount - Desktop, clothespin;

- OS Compatibility - Android 5.0 / Linux 2.6.21 / Windows 10 / Windows 7 / Windows 8 / Windows Vista / Windows XP / macOs 10.10;

- Field of view - 90 °;

- Frame rate per second - 20;

- Connection interface - USB 2.0;

- Additional features - Built-in high-sensitivity microphone Wide angle in FullHD mode;

Moreover, during the work, additional qualities were used, namely a collapsible design and the ability to select a resolution for program (MJPG_1920x1080, MJPG_1280x720, MJPG_640x360).

4.2. Preparing the installation

Due to the fact that, in order to obtain adequate results during the study, it was necessary to make a stable structure, since the cameras should be parallel and be at the same height, the simplest but rigid connection was made between a board with a size of 115x600x20 mm and a shaped wooden corner 300x30x30 mm located in the center of the board, two cameras were fixed from above with an aluminum profile 300x25x3 mm. As a result, the installation looks like the one shown in Fig. 4.2.

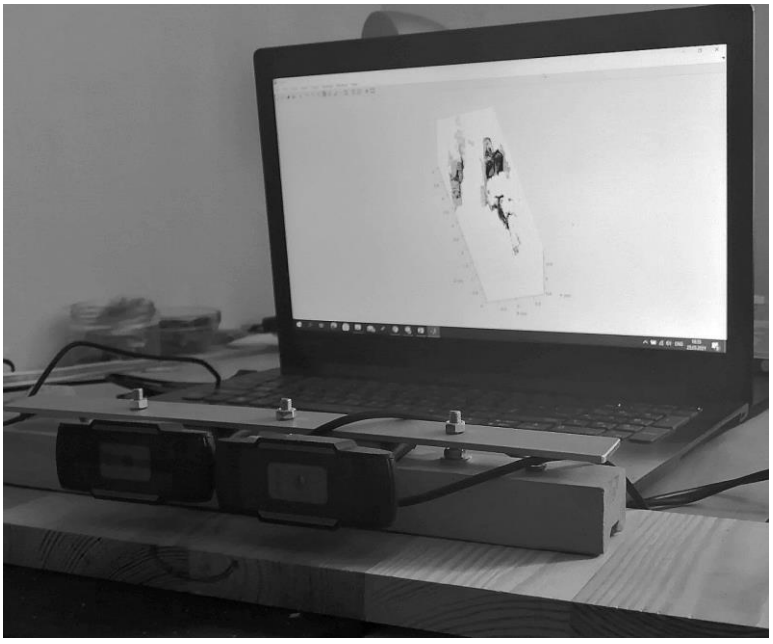

Fig. 4.2. Installation type

After use, it is worth checking the plane of the stand for horizontal position, this can be done using any compass installed on almost every modern phone. It is also worth adding that if the experiment is carried out in a moving plane, then it is necessary to add stable axes of movement (rollers for technical structures), since during shifts there is a risk of displacement relative to the horizon axis, which can significantly affect the effectiveness of the work.

4.3. Preparing cameras for use on software

In order to have a suitable result, the cameras must be located in parallel on the structure, and also along the X, Y axes. For this, the code (Appendix A) was developed based on the MATLAB software, like the entire project as a whole. Step-by-step algorithm for displaying parallel axes from video cameras:

1) Enter into the program the number of inputs with the appropriate equipment;
2) Create an endless image capture;
3) Register pre-showing streaming images;
4) Create a function to enable and disable image capture;
5) Create a cycle for processing images with a delay of half a second, where it is displayed the central axes for the left and right cameras;
6) Manually set the central axes as parallel as possible;
7) Terminate the program.

An example can be seen in Fig. 4.3, as it can be seen, the axes are shown with red lines, and the cameras must be set manually.

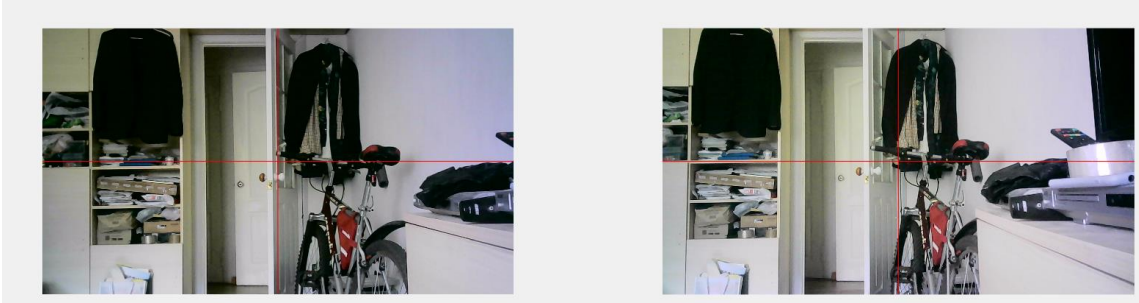

Fig. 4.3. Central axes on two cameras

4.4. Calibration process for two cameras

Calibration is the process of evaluating camera parameters. This means that for future work program will have all the information (specifications or design factors) about the camera to determine the exact relationship between the 3D point in the real world and the corresponding 2D image, the projection (each pixel) in the image captured by the calibrated camera.

Typically, it is needed to get two kinds of parameters for this:

- Internal parameters of the camera / lens system. For example, the focal length, optical center, and radial distortion of the lens;
- External parameters associated with the orientation (rotation and displacement) of the camera relative to some world coordinate system.

4.4.1. Camera calibration methods

The main methods for calibrating a camera are listed below:

1) Template Calibration: With full control over the imaging process, the best way to calibrate is to take multiple images of an object or template of known sizes from different angles. The checkerboard method falls into this very category. Instead of a checkerboard pattern, circular patterns of known sizes can be used;

2) Geometric primitives. Sometimes there are other geometric primitives in the scene, such as straight lines and intersection points, which can also be used for calibration;

3) Based on deep machine learning: When the options for managing image settings are very limited (for example, a single image of a scene), it is still possible to obtain information for camera calibration using Deep Machine Learning.

4.4.2. Calibration Algorithm

In this work, a template calibration method was chosen. The calibration process is explained by the block diagram below (Fig. 4.4).

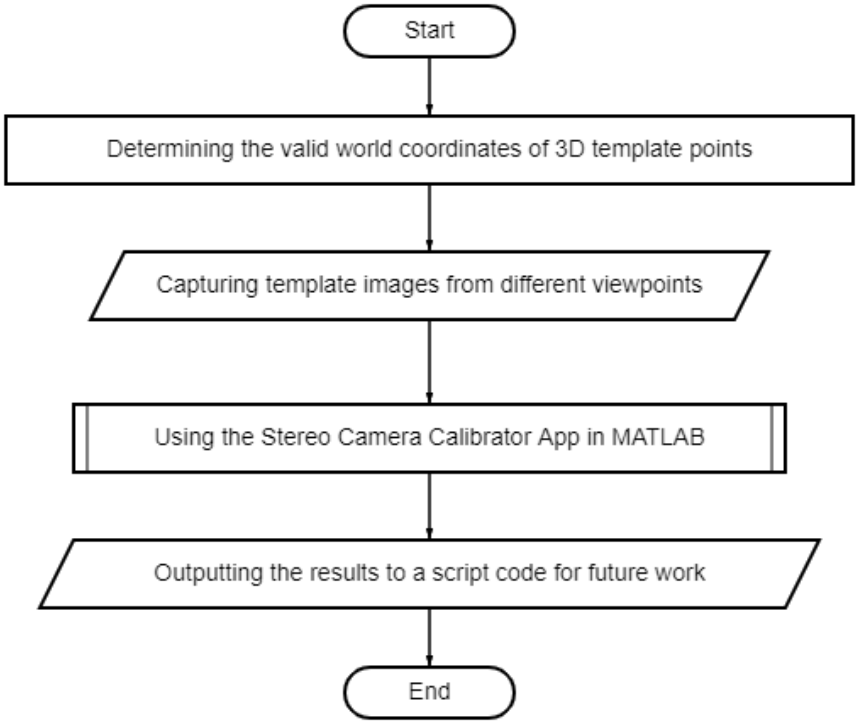

Fig. 4.4. Camera Calibration Block Diagram

1) Determination of the valid world coordinates of 3D template points.

3D points are the corners of the checkerboard cells. Any corner of the board can be selected as the origin of the world coordinate system. The X_w and Y_w axes are along the board, and the Z_w axis is perpendicular to the board. Therefore, all points on the chessboard are on the XY plane (i.e. $Z_w = 0$).

For 3D points, a checkerboard is photographed with known dimensions in different orientations (Fig. 4.5). The world coordinate is linked to the chessboard. Since all corner points lie on a plane, observer can arbitrarily choose Z_w for each point equal to 0. Since the points on the chessboard are located at an equal distance, the coordinates (X_w and Y_w) of each 3D point can be easily determined by taking one point as the origin (0, 0) and determining the remaining point relative to this reference point.

Fig. 4.5 World coordinate system: world coordinates are associated with a checkerboard pattern that is pinned to the board

The checkerboard squares are easily distinguishable in the image and are easy to spot. Moreover, the corners of the checkerboard squares are ideal for localizing them, as it has sharp gradients in two directions. In addition, the corners are at the intersection of the checkerboard lines. All these factors determine the reliability of the search for the corners of staggered squares.

2) Capture template images from different viewpoints

To perform this step of the algorithm, a code (Appendix B) was written, during which the image was read from the left and right cameras, then a cycle was created, in which the final condition was to obtain 50 images. In order to be able to change the position when taking an image, a time delay was added, which can be set to fit your needs and requirements. Thus, two images were taken simultaneously (Fig. 4.6).

Fig. 4.6. Simultaneous shooting of the left and right images

3) Using the Stereo Camera Calibrator App.

This application is built into the MATLAB software and is responsible for calibrating a stereo camera, which can then be used to restore depth from an image, acquire 3D scenes, and so on.

Its interface is designed in such a way that when starting work, person need to add images from the left and right cameras, as well as the exact size of the square of the checkerboard, and the distance is recorded with millimeter precision. After that, it can be analyzed the cases that suit us and add the calculation of radial distortion by 2 and 3 coefficients, add the parameters of the skew and tangential distortion. In case of performed work, two coefficients were chosen, the skew and tangential distortion parameters. The result can be seen in Fig.4.7, Fig.4.8, Fig.4.9, in the process of processing some frames were deleted, as its results negatively influenced the average error per pixel.

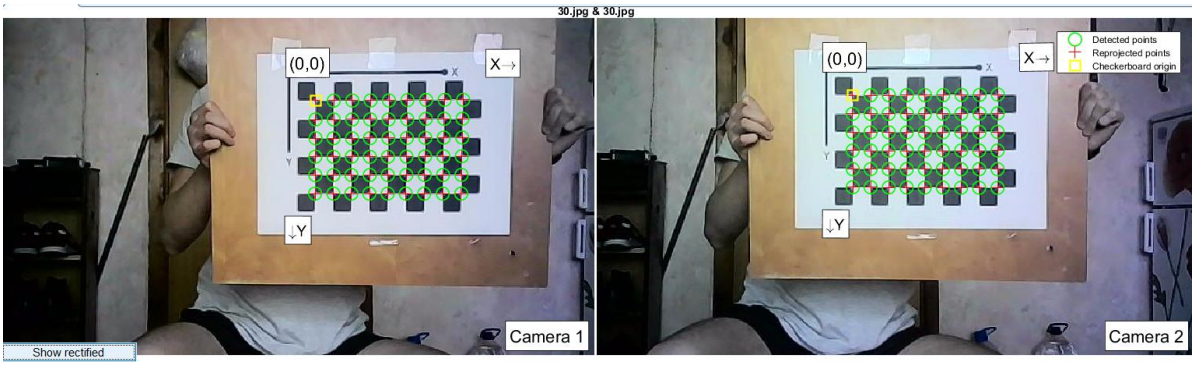

Fig. 4.7. The coincidence of the axes and the starting points of the template

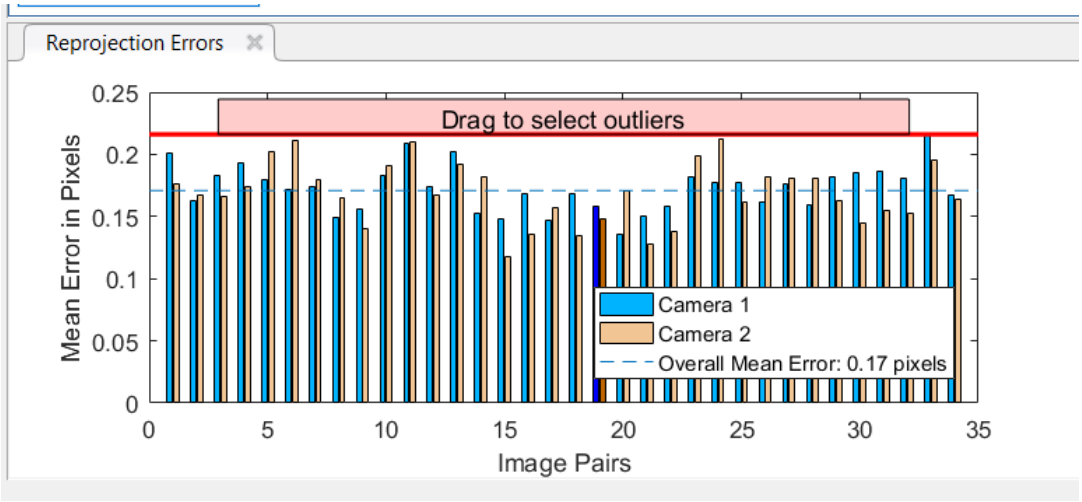

Fig. 4.8. Average error per pixel

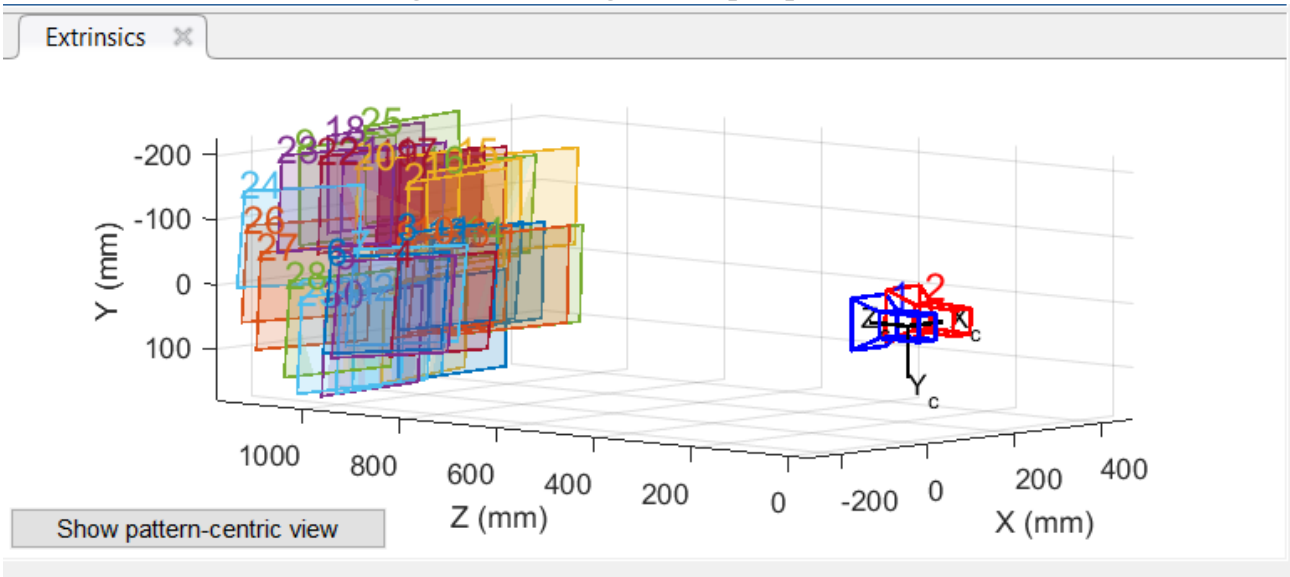

Fig. 4.9. The position of the cameras in space

4) Output of the received data to the script code.

All received data will be added to the script code, which must be saved in advance and called manually, so as not to do it manually, each time the “Load” command is written in the program code, which loads all the data used (Fig. 4.10).

1x1 stereoParameters

Property	Value
CameraParameters1	*1x1 cameraParameters*
CameraParameters2	*1x1 cameraParameters*
RotationOfCamera2	[0.9998 2.0027e-05 -0.0215;-9.9444e-04 0.9990 -0.0453;0.0215 0.0453 0.9987]
TranslationOfCamera2	[-79.8167 4.1308 2.8562]
FundamentalMatrix	[-1.7804e-07 -5.9798e-06 0.0069;2.2335e-06 -6.9684e-06 0.1115;-0.0061 -0.1079 0.7133]
EssentialMatrix	[-0.0890 -3.0402 3.9963;1.1372 -3.6146 79.7778;-4.1314 -79.7308 -3.7015]
MeanReprojectionError	0.1749
NumPatterns	36
WorldPoints	*54x2 double*
WorldUnits	'mm'

Fig. 4.10. Generated stereo parameters in the script code

4.5. Code generation for calculating the depth map

Before creating the code for processing streaming filming, it was necessary to figure out and write an operation algorithm (Fig. 4.11) for at least two images removed from the installation, which was done, and then add the ability to view live objects. Below will be described step by step how this procedure was implemented.

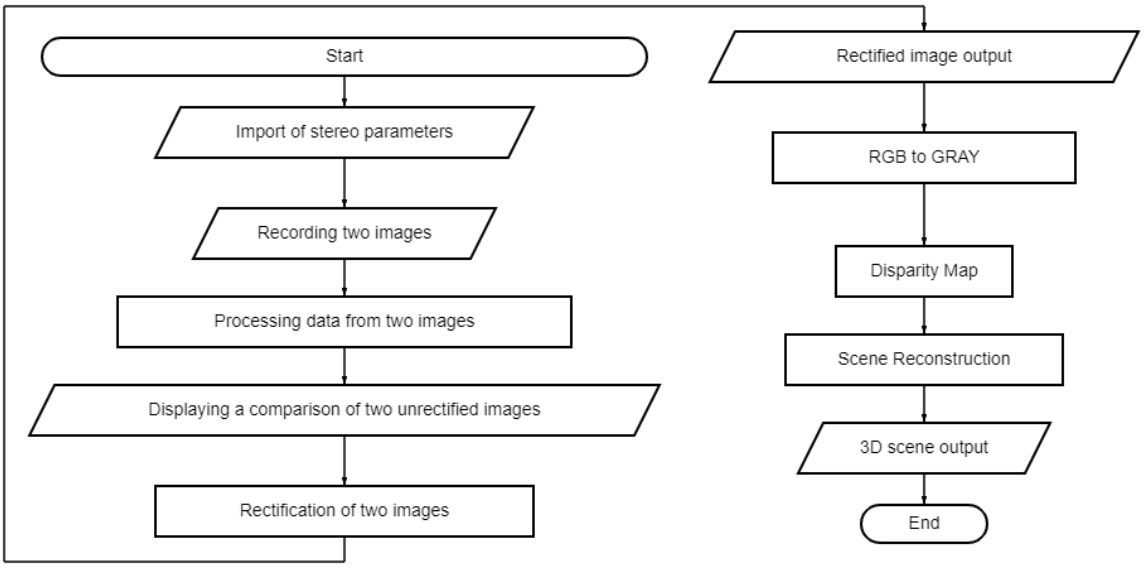

Fig. 4.11. Algorithm of the program

1) As already described earlier, stereo parameters are one of the most important parts of this work and their import into the code directly allows to develop a program for own needs. When importing, all data is transferred and the position of the cameras in space is displayed on the screen, as shown in Fig. 4.9.

2) Then the RGB image is read in uint8 format (it represents an integer from 0 to 255 and takes 1 byte), its size depends on the selected shooting resolution, to get the result it is enough to use an image in 640x360 format, it depends on what resolution want to choose picture. For more correct conclusions, it is needed to smooth the image, and this can be done thanks to the Gaussian filter, which uses the normal distribution to calculate the transformation that is applied to each pixel of the two images. Each element of the image corresponds to a number called a weighting factor. The sum of all the weighting factors constitutes the weighting function. To unambiguously determine the central element, the window size must be odd. Thus, a 3x3 convolutional window was chosen. The convolution kernel allows to enhance or attenuate image components. The matrix moves across the image, while the weighting function remains unchanged during the movement.

3) The output of two unrectified images (Fig. 4.12) shows an inconsistency in Y and proves that this must be corrected in subsequent steps.

Fig. 4.12. Unrectified Stereo Image

As it is shown in the image, this is solved by overlaying one image on top of another, using muted halftones. To make sure of the above, it will be enlarged a certain area (Fig. 4.13) of this picture.

Figure 4.13. Height difference

4) In order to get rid of this problem, rectification of stereo images is carried out. Image rectification (Fig. 4.14) is the transfer of two image planes into one plane so that all epipolar lines are parallel to the abscissa axis and the corresponding epipolar lines in both images have the same ordinates. After that, undistorted and corrected versions of the input images are returned, taking into account the previously prepared stereo parameters, taking into account bevels, transformations, and the like, conventionally this sub-clause solves the issue of failures in the location of cameras and makes their focal positions parallel.

Fig. 4.14. Rectified image frames

It is worth noting that thanks to this function, the 3D effect is already created, if the viewer put on special 3D glasses, viewer get the effect of the three-dimensional vision of the picture, as if watching a movie in a cinema.

5) As seen in Fig. 4.14, the conclusion is made with some cropping of the photo, this is due to the fact that when correcting images, a suitable position and angles of rotation of the camera are sought, while black zones appear (Fig. 4.15) or holes in the plane, which negatively affect further work, because of this, the image is cropped to the size that is valid (valid parts).

Fig. 4.15. Uncropped rectified image

6) The conversion of RGB to Gray (grayscale) is made due to the fact that the disparity map is not able to work with a color image, but works on grayscale, that is, this is the color mode of images that are displayed in grayscale, placed in a table as standards of white brightness. Most often, a stepped image of a uniform series of optical densities of neutral-gray fields is used.

7) This stage is decisive, since when displaying the Disparity Map, a simple colored analogue of the location of objects at a distance is shown. That is, if the object is in any of the positions, it is highlighted in one color, in accordance with the distance, that is, the closer, the lighter its shade will be (Fig. 4.16).

Fig. 4.16. Disparity map in gray scale

It can be noticed that there is some disparity with reality, but this may arise due to the monotony of the color of the object itself, or because of the glare of light on the camera, it was also found during experiments that when transparent bodies or with a mirror base appear in areas, parts of them may not be displayed due to their structure. To make it more pleasant and visual for a person to examine the scenario, it was decided to add a color palette with the dimension [0 64] (Fig. 4.17). It is important to note that the number of shades can be changed manually, or not specified at all, in this case the minimum and maximum values are accepted depending on the type of the recorded format, for example for uint16 this value is [0 65535].

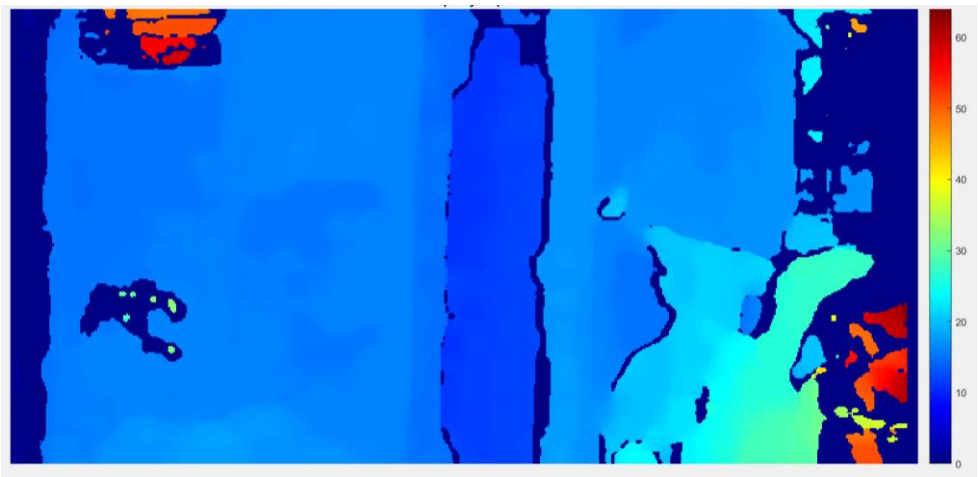

Fig. 4.17. Disparity Map in color

8) Scene Reconstruction is carried out by binding stereo parameters to the disparity map, that is, if a normal disparity map model is received, it is sent to create a scene, since all world coordinates of points stored in the image are already stored in it. In this case, the parameter input must be the same as that used for correction, taking into account all corrections of stereo images that were applied before the corresponding disparity maps. After that, having received the data with the depth of each point, data must be rewritten in meters, since the initial data is recorded in millimeters, for this all internal values are divided by 1000. Then it can be displayed the 3D scene on the screen, but it will not be very understandable for a person, due to the fact that only data from the Disparity Map array is sent there, and, as mentioned earlier, a picture is displayed there only with halftones and to make the project realistic, values from one of the color images are added to the code. It is also worth adding garbage collection when displaying on the screen, since it was interested in a scene no more than 2 meters in width and height and 5 in length, all other values were simply thrown out of consideration, it was a configurable parameter and applied under personal conditions.

9) Displaying is performed by a simple function of filling volumetric figures, with which it can be done various manipulations such as increasing rotation and capturing a point. When capturing, the cameras should be horizontal and not subject to focal changes during the process. Examples of several rooms are Fig. 4.18, Fig. 4.19, Fig. 4.20, Fig. 4.21.

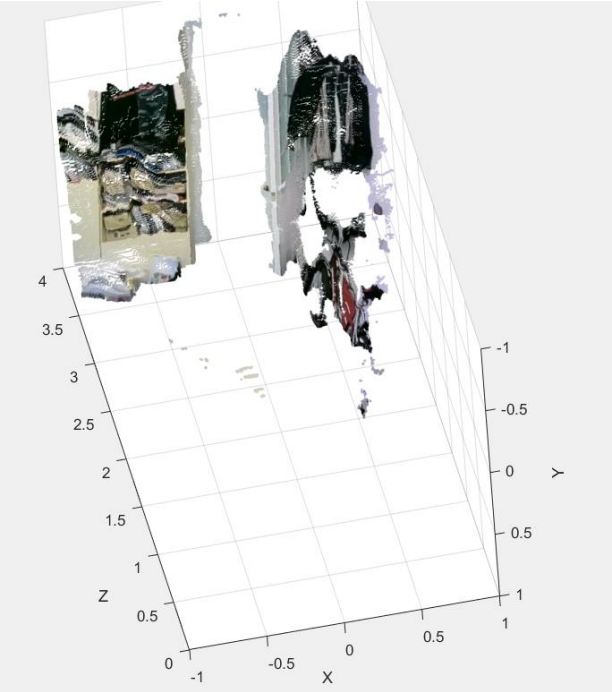

Fig. 4.18. Example A

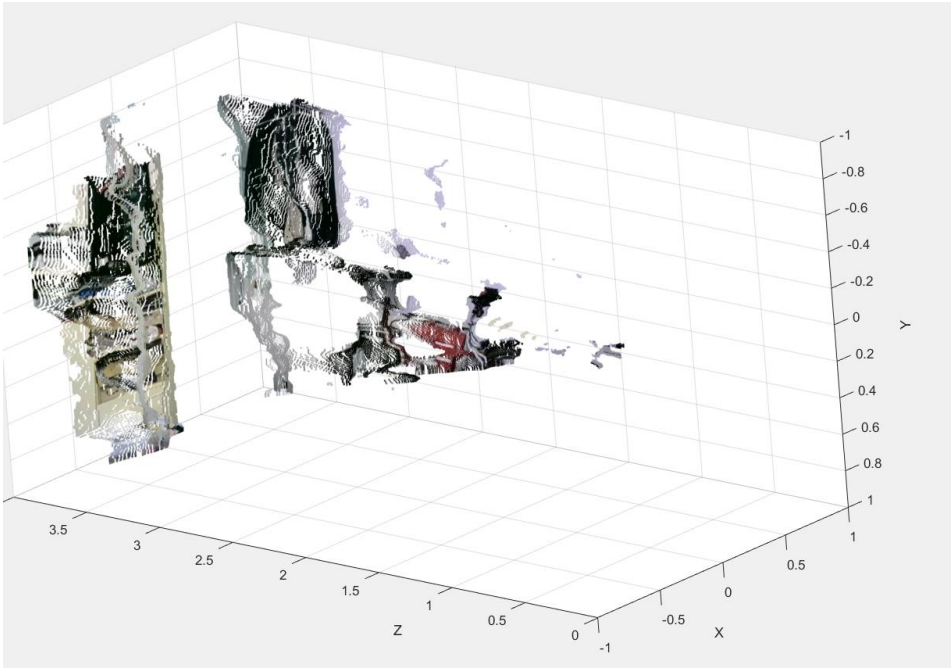

Fig. 4.19. Example A

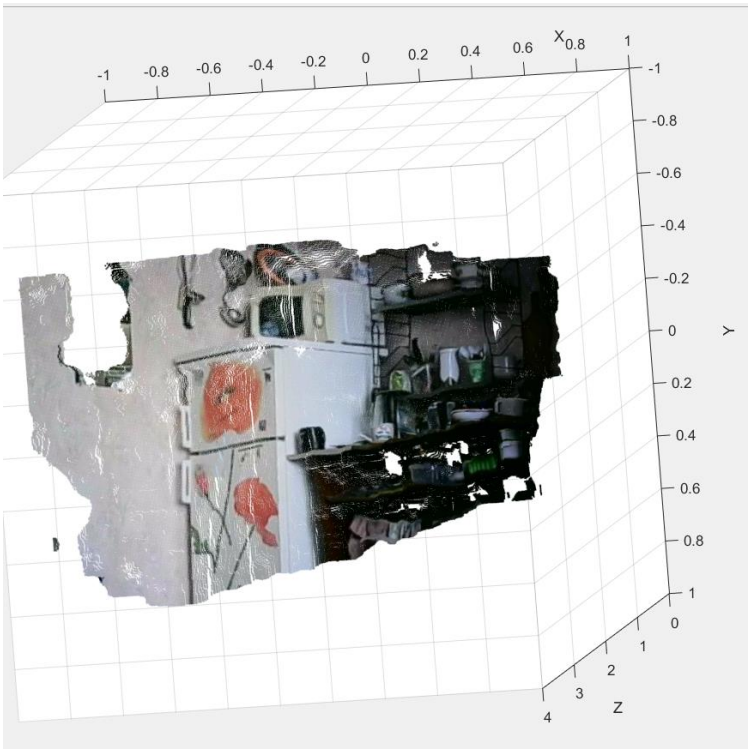

Fig. 4.20. Example B

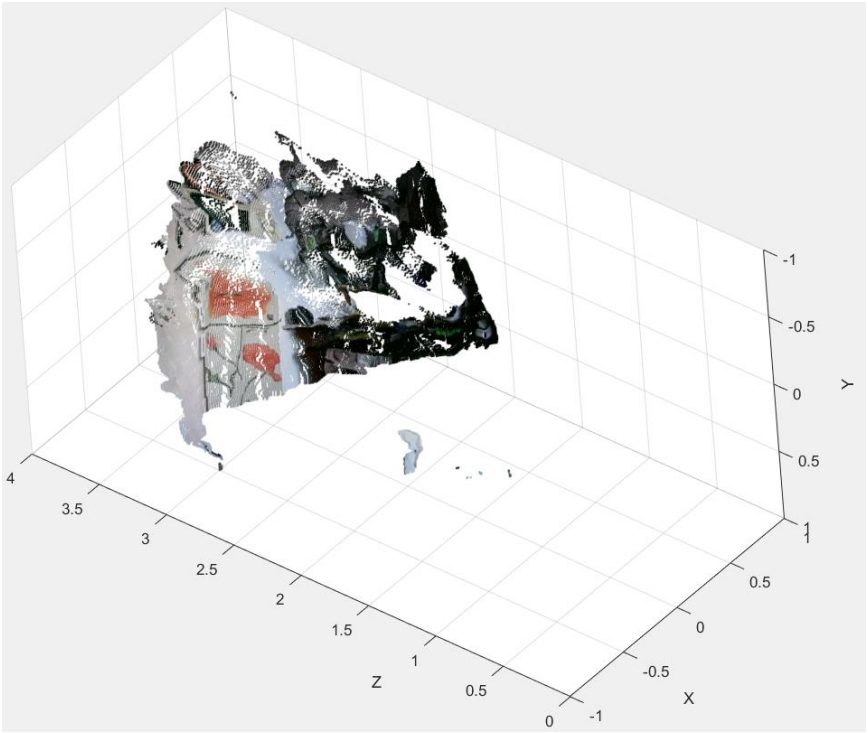

Fig. 4.21. Example B

From these scenes, it can be easily found out the approximate distance to something, see a three-dimensional picture with cutouts in those places where the

bodies cover each other or are not in the field of view, as can be seen when the scene rotates.

Streaming is implemented, in the same way as the described algorithm for images, only it works on the principle of rewriting images from the left and right cameras through a minimum time delay and the output is only a 3D scene, since this process is in a loop and when calling any other it is called a large number of times, which is uncomfortable for viewing and understanding, rotation of the scene is also unacceptable, since the figure is built many times and re-displays to the starting position of the XYZ axes (Fig. 4.22).

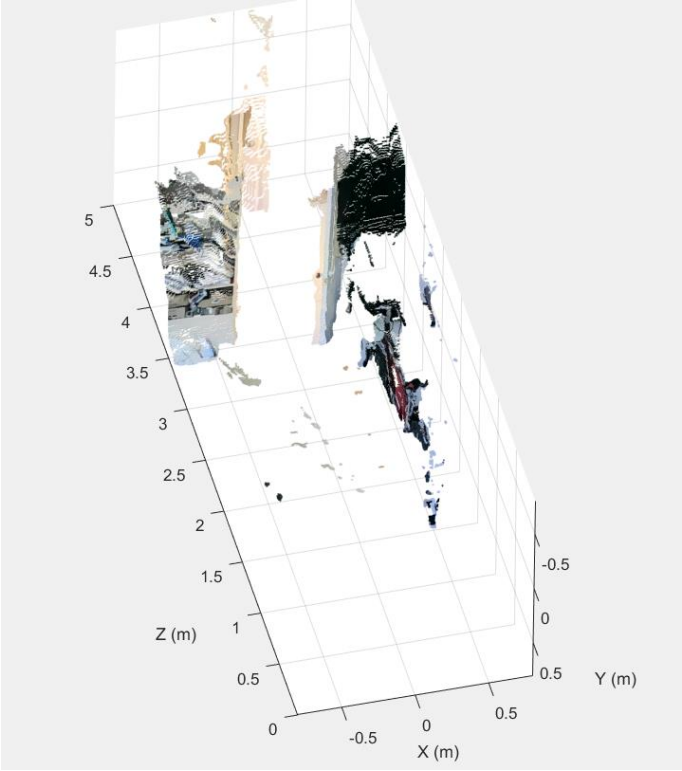

Fig. 4.22. Streaming 3D scene

If necessary, the work of the cycle can be stopped and the data that was recorded at the last moment is saved for future work with them.

4.6. Analysis estimation of the errors and accuracy

To confirm the good performance of the installation, an analysis was carried out taking into account the accurate calculation of the range to objects. Initially, the body was placed at a different distance with a range of 20 centimeters, and moved

within a range of 4 meters. Since the work was done in a room environment, a distance of more than 5 meters was not of interest. The result of the experiment is presented in Table 4.1.

Table 4.1. Accuracy check against real and measured range with depth map

Undisturbed environment			
Real distance, m	Distance measured by the depth map, m	Absolute Error, m	Relative Error, %
1	1,05	0,05	5,00
1,2	1,27	0,07	5,83
1,4	1,35	-0,05	3,57
1,6	1,52	-0,08	5,00
1,8	1,86	0,06	3,33
2	2,04	0,04	2,00
2,2	2,26	0,06	2,73
2,4	2,43	0,03	1,25
2,6	2,58	-0,02	0,77
2,8	2,84	0,04	1,43
3	3,04	0,04	1,33
3,2	3,18	-0,02	0,63
3,4	3,41	0,01	0,29
3,6	3,64	0,04	1,11
3,8	3,78	-0,02	0,53
4	4,01	0,01	0,25

As is shown, real data and data taken from the program are presented, after which it was possible to calculate the absolute and relative error of the work. On the graph (Fig. 4.23) can be seen the ratio of values.

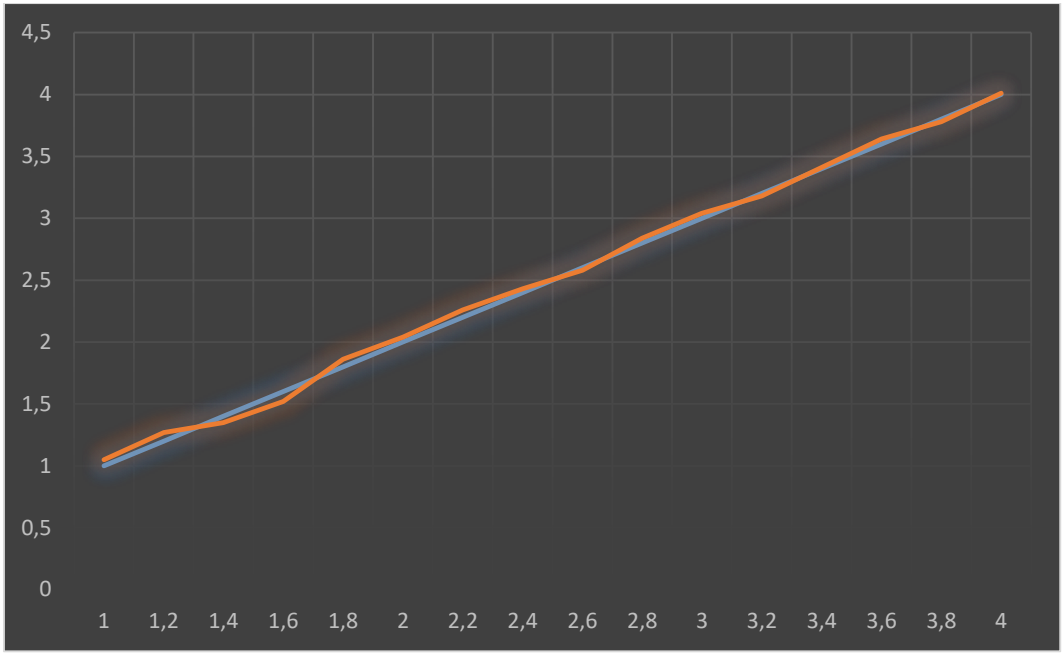

Fig. 4.23. Comparison of data

Thus, the model can be called adequate, since the error does not exceed 5 percent, which is for devices that are designed to measure the distance to an object.

Also, the accuracy results can vary depending on where the device is located, as the environment can affect for a number of reasons, such as light, surface reflection, pressure, etc.

4.7. Computing power consumption

Nowadays, an important role is played by how quickly the program can work, because of this developer need to know the costs and what equipment is needed for high-quality work without sagging. It is important that the program works without failures and sagging, since it is mainly intended for live filming, and during this, a strong load on the processor, video card and required memory may occur.

The block diagram of the output device of the streaming 3D scene has the form shown in Fig. 4.24.

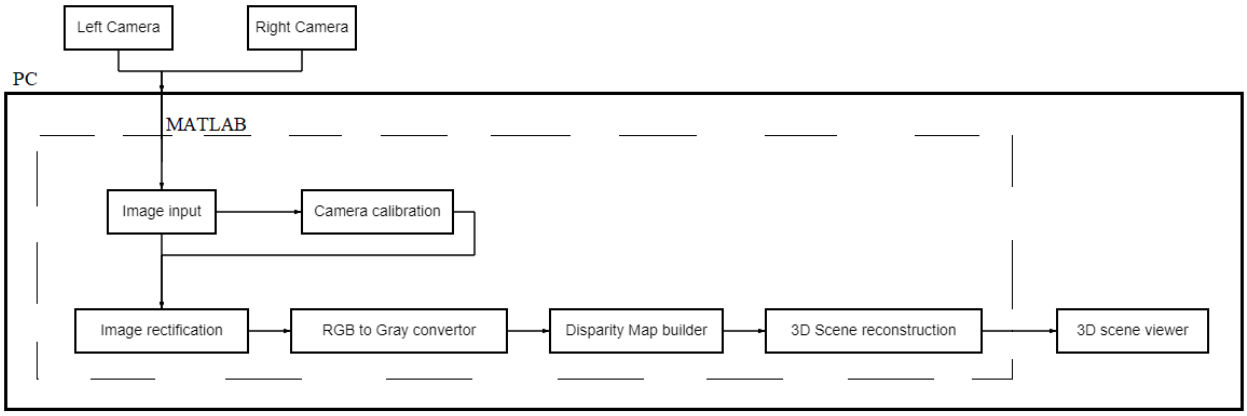

Fig. 4.24. Block diagram of the output device of a streaming 3D scene

In order to find out how fast the program can work, for a start, the characteristics of my PC (Table 4.2) and the cost of the used memory will be presented. PC that was used for work doesn't have the best system requirements, but even that is enough.

Table 4.2. Characteristics of PC

PC name	Lenovo Ideapad 320 15i-KB
CPU	Intel Core i3-7130U CPU 2,7 GHz
Number of Cores	2
Logical processors	4
Installed RAM	12 Гб
Video adapters	Intel(R) HD Graphics 620 NVIDIA GeForce 940MX

Then, an experiment was carried out with the load on the system when processing one image, during which it was proved that the load on the CPU did not exceed 20%, and the processing time did not exceed 3 seconds, this number may also seem too large, but just think, there is a search and transportation of the image from the PC disk to the program, after which it is already converted with various display methods, for a more illustrative example. From my point of view, these results are quite acceptable for comfortable work with one image. By the way,

memory costs were only 900 MB, which is very undemanding for modern demanding programs, which certainly gives me a big plus.

Now let's move on directly to the work of streaming video. It is clear that the memory costs will definitely be much more than when working with one image, but there is a significant plus, in the form of the fact that the speed of the loop will be much faster than the entire passage of the program code (Appendix C) for one scene. This arises due to the fact that there is no need to call and show unnecessary information, and in the end, the same window is simply overwritten.

In order to find out the exact time of passing one cycle, the tic toc function (Fig. 4.25) in MATLAB was used, which works on the clock principle, tic - start, toc - end of timer.

```
Elapsed time is 19.528022 seconds.
Elapsed time is 19.943868 seconds.
Elapsed time is 20.364600 seconds.
Elapsed time is 20.785838 seconds.
Elapsed time is 21.225324 seconds.
Elapsed time is 21.696200 seconds.
Elapsed time is 22.144253 seconds.
Elapsed time is 22.715397 seconds.
```

Fig. 4.25. Tic toc function with capturing the cycle time of a stream program

After the experiment, it can be proved that the scene is rewritten about twice per second, this is a very good result, since the camera itself does not exceed 4 frames per second, and taking into account the processing, it achieved two frames per second.

Let's move on to the required memory, the CPU began to load by 10% more than during normal operation, but the surprise was that the spent memory remained at the same level. This is a satisfactory result, but since the system has not been used on other devices, personally cannot claim that working on other PCs will not be more stressful.

CONCLUSION

After performed work, it was proved that when using two web cameras and the MATLAB software environment, it can be created an effective stereo pair for analyzing the environment and building 3D scenes.

The main disadvantages of the method are its adjustment, namely calibration, obtaining an adequate model, etc. Difficulties also arose with the creation of a stable stand, which would make it possible, when the cameras were shifted, to make adjustments to their position relative to the axes.

In addition, nuances were noticed regarding different areas, projective planes and display on the graph itself, since the light is not constant, some glare, transparency of bodies, or vice versa - specularity are possible. An interesting fact was noticed with mirrors, which were cut out of consideration due to the fact that the cameras could not understand and match positions with the real world and objects in the mirror did not have a spatial position.

An analysis of the range accuracy showed that the system does not exceed a relative error of more than 5%, which is a fairly acceptable result even for rangefinder tasks using other methods, and if take into account that the installation is made of cheap materials and equipment, then the result can be considered quite successful.

REFERENCES

1. Где сегодня применяется компьютерное зрение. *Центр Международной Торговли*. URL: https://corp.wtcmoscow.ru/services/international-partnership/actual/gde-segodnya-primenyaetsya-kompyuternoe-zrenie/ (date of access: 19.05.2021).

2. Что такое компьютерное зрение и где его применяют | РБК Тренды. *РБКТренды*.URL: https://trends.rbc.ru/trends/industry/5f1f007e9a794756fafbfa83 (date of access: 16.05.2021).

3. Ализар А. Генерация 3D-моделей по фотографиям. *Все публикации подряд / Хабр*. URL: https://habr.com/ru/post/64080/ (date of access: 17.05.2021).

4. Махмутова Г.Э. Моделирование и исследование процессов стереовидения. МЭИ 2012.

5. Волосов Д. С. Фотографическая оптика. М., «Искусство», 1971.

6. Как вычислить расстояние до объекта по фотографии. *Ретротехника, самоделки и борьба с идиотизмом – LiveJournal*. URL: https://bootsector.livejournal.com/43436.html (date of access: 28.04.2021).

7. Технология NanoLOC. *Самый информированный сервер микроэлектроника, описания - rs232, rs 232, микросхемы, hd44780, atmel, ацп, цап, irda, микроконтроллер.* URL: http://www.gaw.ru/html.cgi/txt/doc/Wireless/nanonet/index_loc.htm (date of access: 29.04.2021).

8. Точность определения расстояний с помощью технологии nanoLOC - Журнал Беспроводные технологии. *Журнал Беспроводные технологии*. URL: https://wireless-e.ru/rtls/receive-signal-strength-indication/ (date of access: 28.04.2021).

9. “Как устроены и работают инфракрасные датчики движения” Сайт для электриков - советы, примеры, схемы. *Сайт для электриков - советы, примеры, схемы.* URL: http://electrik.info/main/automation/917-kak-ustroeny-i-rabotayut-infrakrasnye-datchiki-dvizheniya.html (date of access: 25.04.2021).

10. Kinect.*ru.freejournal.org*. URL: https://ru.freejournal.org/1818924/1/kinect.html (date of access: 25.04.2021).

11. Лидар. Применение технологии LiDAR. Карты и беспилотные автомобили | Gistroy. *Настольный лазерный гравер. Купить мини лазерный гравер с ЧПУ | Gistroy*. URL: https://gistroy.ru/article/lidar/ (date of access: 08.05.2021).

12. Датчик расстояния HC-SR04 - ультразвуковой модуль Ардуино. *ArduinoMaster все об Ардуино*. URL: https://arduinomaster.ru/datchiki-arduino/ultrazvukovoj-dalnomer-hc-sr04/ (date of access: 09.05.2021).

13. Технология построения 3D-моделей объектов по набору изображений URL: http://masters.donntu.org/2012/fimm/solovyov/library/article7.htm (date of access: 10.05.2021).

14. LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998), "Gradien-based learning applied to document recognition". *Proceedings of the IEEE*, Vol. 86, Issue 11, Nov. 1998, pp. 2278- 2324.

15. Krizhevsky, A., Sutskever, I. and Hinton, G.E. (2012), "ImageNet Classification with Deep Convolutional Neural Networks". *NIPS'12 Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097-1105.

16. Карта глубины – Викиконспекты. *Northern Eurasia Contests*. URL: https://neerc.ifmo.ru/wiki/index.php?title=%D0%9A%D0%B0%D1%80%D1%82%D0%B0_%D0%B3%D0%BB%D1%83%D0%B1%D0%B8%D0%BD%D1%8B (date of access: 10.05.2021).

17. Бакулев П.А. Радиолокационные системы: *Учебник для вузов.: Радиотехника*, 2004.

18. Kurakin A. Основы стереозрения. *Все публикации подряд / Хабр*. URL: https://habr.com/ru/post/130300/ (date of access: 10.05.2021).

19. Стереоскопическое компьютерное зрение URL: https://compress.ru/article.aspx?id=9529 (date of access: 10.05.2021).

20. Гонсалес Р., Вудс Р. Цифровая обработка изображений. – М.: Техносфера, 2005.

21. Buades, A., B. Coll, and J.M. Morel, "A review of image denoising algorithms, with a new one," *SIAM Multiscale Modeling and Simulation*, vol. 4, pp. 490.530, 2005.

22. Грузман И.С, В.С. Киричук, В.П. Косых, Г.И. Перетягин, А.А.Спектор. Цифровая обработка изображений в информационных системах: Учебное пособие. — Новосибирск: Изд-во НГТУ, 2000. — 168.

23. Апальков И.В., Хрящев В.В. Удаление шума из изображений на основе нелинейных алгоритмов с использованием ранговой статистики. — Ярославский государственный университет, 2007.

APPENDIX

Appendix A. The program created for the location of the central axes on the camera image in the Matlab environment

```
vid1    =    videoinput('winvideo',    1,    'MJPG_1920x1080');
%second value is the port number (worth clarifying), the
third is the resolution (left camera)
vid2 = videoinput('winvideo', 3, 'MJPG_1920x1080');     %The
right camera input

vid1.TriggerRepeat = Inf;   %Continuous image capture
vid2.TriggerRepeat = Inf;

preview(vid1);       % preview from cameras
preview(vid2);

 global LOOP_RUNNING;
LOOP_RUNNING = true;

while (LOOP_RUNNING)

     imgR = getsnapshot(vid2);
     imgL = getsnapshot(vid1);

     % creating centerlines

     subplot(1,2,1), imshow(imgL)
     hold on;
     line(  [  0,  1920  ],  [540,  540  ],  'Color',  'red',
     'LineStyle','-');
     line(  [960,  960],  [0,  1080]  ,  'Color',  'red',
     'LineStyle','-');
     subplot(1,2,2), imshow(imgR)
     hold on;
     line(  [  0,  1920  ],  [540,  540  ],  'Color',  'red',
     'LineStyle','-');
     line(  [960,  960],  [0,  1080]  ,  'Color',  'red',
     'LineStyle','-');

     pause(0.5);
end


stop (vid1);
stop (vid2);
```

Appendix B. Program designed to shoot 50 pairs of stereo images

```
vid1 = videoinput('winvideo', 1, 'MJPG_640x360');
vid2 = videoinput('winvideo', 3, 'MJPG_640x360');

preview(vid1);
preview(vid2);

pause(5);

imgL = getsnapshot (vid1);
imgR = getsnapshot (vid2);

stop (vid1);
stop (vid2);

n=2;                    % pause time between shots
    tL_1=['E:\Left\'];        % save folders link
    tL_3 =['.jpg'];

    tR_1=['E:\Right\'];
    tR_3=['.jpg'];
for i = 1:50
    imgL = getsnapshot (vid1);
    imgR = getsnapshot(vid2);

    tR_2 = int2str(i);
    tL_2 = int2str(i);
    tL = strcat(tL_1,tL_2,tL_3);
    tR = strcat(tR_1,tR_2,tR_3);
    imwrite (imgR,tR ,'jpg')
    imwrite (imgL, tL, 'jpg')
    pause(n);

end
```

Appendix C. 3D scene output program

```
leftCam = imaq.VideoDevice('winvideo', 1, 'MJPG_640x360');
rightCam = imaq.VideoDevice('winvideo', 3, 'MJPG_640x360');

leftCam.ReturnedDataType = 'uint8';

rightCam.ReturnedDataType = 'uint8';

if ~exist ('stereoParams', 'var')
    load stereOB4
end

ax = axes;
maxDepth = 5;

while true
    imageLeft = step(leftCam);
    imageRight = step(rightCam);

    [J1, J2] = rectifyStereoImages(imageLeft, imageRight,
stereoParams);

    disp       =    disparity    (rgb2gray(J1),rgb2gray(J2),
'DisparityRange', [0, 64] );

    pointCloud    =    reconstructScene(disp,    stereoParams)
./1000;

    z = pointCloud (:, :, 3);
    z (z < 0) = NaN;
    z (z > maxDepth) = NaN;
    pointCloud (:, :, 3) = z;

    if ~ishandle(ax)
        break;
    else
        pcshow(pointCloud, J1, 'VerticalAxis', 'Y', ...
            'VerticalAxisDir', 'Down', 'Parent', ax);

        xlabel('X (m)');
        ylabel('Y (m)');
        zlabel('Z (m)');

        xlim (ax, [-.8, .8]);
        ylim (ax, [-.8, .8]);
        zlim(ax, [0, maxDepth]);
        daspect(ax,'manual');
        pbaspect (ax, 'manual');
```

```
        drawnow;

    end
end

release (leftCam);
    release (rightCam);
```