INSTITUTE OF ENGINEERING
PASHCHIMANCHAL CAMPUS
DEPARTMENT OF ELECTRONICS AND COMPUTER ENGINEERING

A Project Proposal On

ADVANCEMENTS IN DEPTH ESTIMATION : TECHNIQUES AND APPLICATION

Presenters:

Avishek Poudel[WRC077BEI009]
Binayak Shrestha [WRC077BEI012]
Prakanda Bhandari [WRC077BEI030]
Pratham Adhikari [WRC077BEI032]

Contents

- Introduction
- Problem Statement
- Objectives
- Application
- Literature Review
- Methodology
- Time Plan and Cost Estimation
- Expected Outcomes
- References

Introduction

- Depth estimation is crucial for understanding and interacting with 3D environments in computer vision.
- Traditional depth estimation methods include TOF (Time-of-Flight) sensors and LIDAR (Light Detection and Ranging).
- TOF sensors and LIDAR measure the time taken by an infrared light pulse or laser pulse to travel to an object and back to calculate distance.
- Traditional methods have limitations such as high cost, large size, high power consumption, and complexity.

Introduction

- Recent advancements in monocular depth estimation leverage machine learning and computer vision techniques.
- Deep learning models, such as Convolutional Neural Networks (CNNs), are trained on large datasets to predict depth maps from single images.
- These models achieve high accuracy by learning intricate patterns and cues that signify depth.
- Monocular depth estimation allows for the generation of dense and accurate depth maps from RGB images.

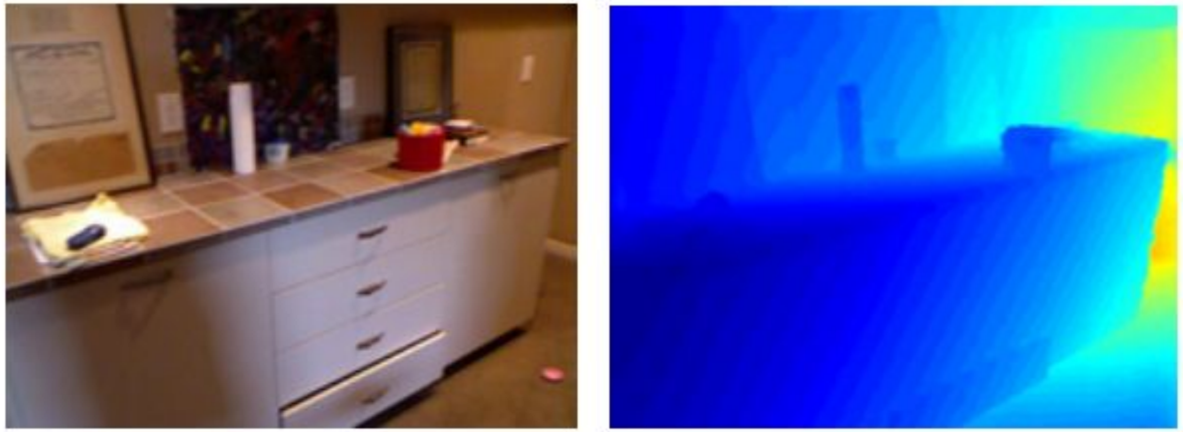


Fig 1: Image and Depth map

Problem Statement

- Accurate depth estimation is essential for various applications in computer vision and robotics.
- TOF and LIDAR provide reliable depth measurements but have significant drawbacks High cost ,Substantial power consumption, increased system complexity.
- Depth estimation from RGB images offers a more accessible and cost-effective alternative. This approach infers depth from visual cues such as texture, shading, perspective, and motion.
- Interpreting these visual cues can be ambiguous and complex.

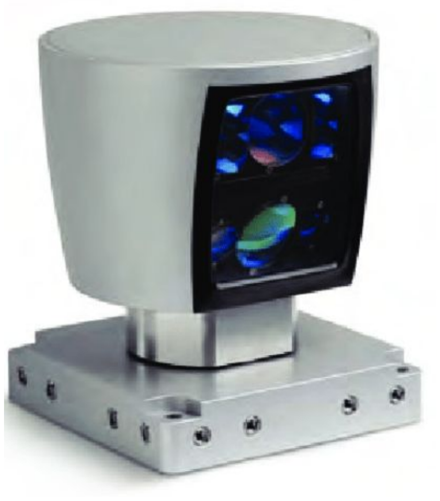


Fig 2:LIDAR

Objectives

The objectives of the project is :

- To design and develop an accurate depth estimation system using RGB images.

Application

- The major application of the project are:
- Enhances the ability of robots to understand and navigate human environments safely and effectively.
- Allows systems to detect and avoid obstacles in real-time, preventing collisions and enhancing navigation.
- Provides accurate 3D maps of the environment, crucial for navigation and situational awareness.
- Improved depth perception in AR/VR system and Realistic object interaction and manipulation and enhances immersive experiences by providing accurate spatial information and enabling precise tracking of virtual objects.

Application

- For 3D reconstructions and modeling, it facilitates the creation of detailed and accurate 3D models for various applications.
- Enhances medical imaging by surgical planning, navigation and diagnosis by providing detailed 3D views of anatomical structures, aiding in precise surgical planning and diagnosis.
- In photography and cinematography, depth maps can improve the quality of images and videos by enabling advanced effects and accurate focus.
- Security and surveillance with enhanced facial recognition and biometrics.
- In telepresence and teleconferencing, it can provide a more immersive and interactive experience by accurately capturing and rendering 3D environments.

Literature Review

Authors	Title	Findings
D. Eigen, C. Puhrsch, and R. Fergus	Depth map prediction from a single image using a multi-scale deep network	Depth estimation from a single 2D color image through a deep neural network. It employed two deep network stacks: one that makes a coarse global prediction based on the entire image and another that refines this prediction locally.
G. Liu, G. Jiang, R. Xiong, and Y. Ou	Binocular depth estimation using convolutional neural network with siamese branches	They introduced an approach of binocular depth estimation method based on deep learning. A new convolutional neural network is designed, which consists of two sub-networks. The first subnetwork is a deep network with Siamese branches and 3D convolutional layer, it learns parallax and global information and generates a global depth estimation

Literature Review

Authors	Title	Findings
		result in low resolution. The second is a fully convolutional deep network, which reconstructions the depth map to original resolution. The two sub-networks are connected by a pool pyramid.
C. Godard, O. Mac Aodha, and G. J. Brostow	Unsupervised monocular depth estimation with left-right consistency	It enables the convolutional neural network to learn to perform single image depth estimation, despite the absence of ground truth depth data. The model uses bilinear sampling to generate images , resulting in a fully (sub-)differentiable training loss. A fully convolutional deep neural network, by posing monocular depth estimation as an image reconstruction problem can solve the disparity field

Literature Review

Authors	Title	Findings
		without requiring ground truth depth. It includes a left right consistency check to improve the quality of synthesised depth images.
S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari	Robust monocular depth estimation under challenging conditions	Use of synthetic data to train the model for handling adverse weather conditions like rain and night using a method called md4all. Md4all utilise existing successful depth estimation methods for ideal conditions. First it generates complex samples corresponding to normal training ones. They trained the model by guiding it self- or full-supervision by feeding the generated samples and computing the standard losses on the corresponding original images. Enables a single model to recover information across diverse conditions without modifications at inference time.

Literature Review

Authors	Title	Findings
H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao,	Deep ordinal regression network for monocular depth estimation	A SID strategy to discretize depth and recast depth network learning as an ordinal regression problem. By training the network using an ordinary regression loss, it achieves much higher accuracy and faster convergence. It adopts a multi-scale network structure which avoids unnecessary spatial pooling and captures multi-scale information in parallel.
D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze	Fastdepth: Fast monocular depth estimation on embedded systems,	An efficient and lightweight encoder-decoder network architecture and apply network pruning to further reduce computational complexity and latency. It demonstrates that it is possible to achieve similar accuracy as prior work on depth estimation, but at inference speeds that are an order of magnitude faster. Addresses the problem of slow depth estimation on embedded systems.

Literature Review

Authors	Title	Findings
R. Furukawa, R. Sagawa, and H. Kawasaki	Depth estimation using structured light flow–analysis of projected pattern flow on an object's surface	Minimum two light flows, which are retrieved from two projected patterns on the object, are required for depth estimation. To retrieve two light flows at the same time, two sets of parallel line patterns are illuminated from two video projectors and the size of motion blur of each line is precisely measured. By analyzing the light flows, i.e. lengths of the blurs, scene depth information is estimated.
C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow	Digging into self supervised monocular depth estimation	A simple model with minimum reprojection loss, designed to robustly handle occlusions, a full-resolution multi-scale sampling method that reduces visual artifacts and an auto-masking loss to ignore training pixels that violate camera motion assumptions.

Methodology

- The project follows an Iterative Development Process, allowing for continuous refinement and adaptation based on learning from each phase of the project.

- Different statistical and mathematical tools are used to establish clear decision points throughout the project lifecycle and define key artifacts (both input and output) to guide the project's progression and ensure alignment with project goals.

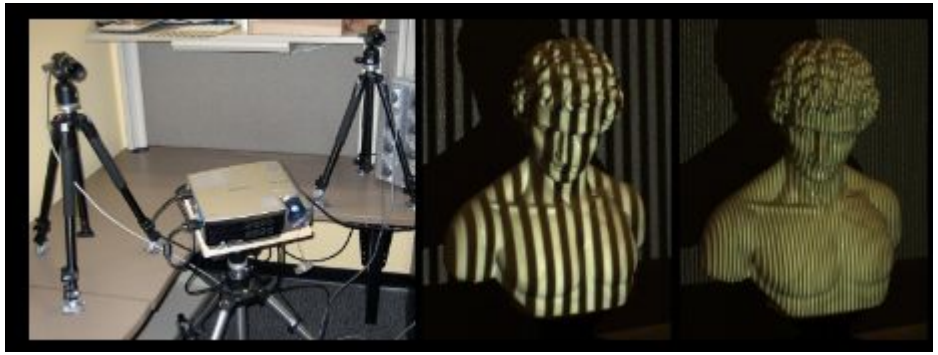


Fig 3: Structural Light Illumination

Methodology

Hardware design

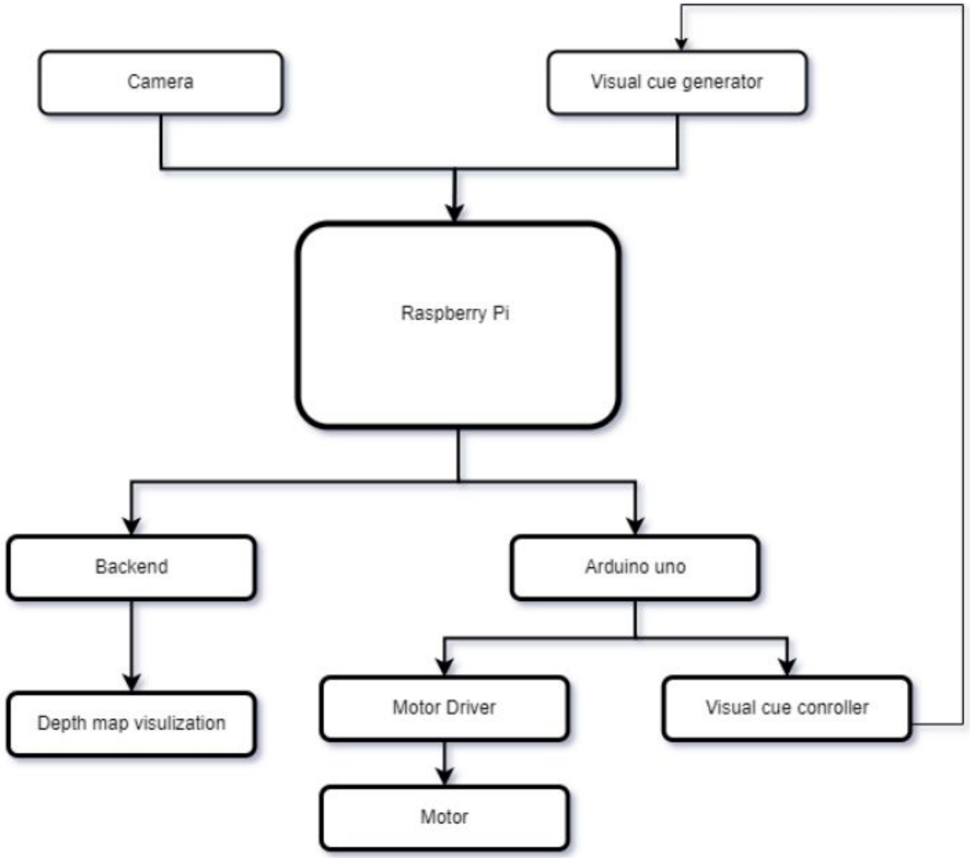


Fig 4: Hardware block diagram

Methodology

Hardware design

- The project uses a Raspberry Pi as the brain.
- Visual Cue generator generates cues that help in understanding how far away objects are.
- An Arduino Uno microcontroller steps in to control the Visual cue generator, and motor driver, which powers the motor.
- The Raspberry Pi processes images captured by camera and sends them to the backend through Wi-Fi for visualization.

Methodology

Hardware design

Fig 5: Arduino Nano

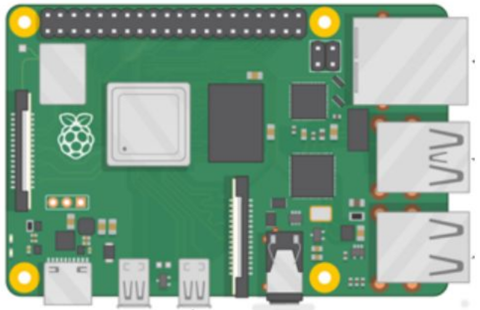


Fig 6: Raspberry Pi

Fig 7: Motor

Fig 8: Servo Motor

Methodology

Software Design

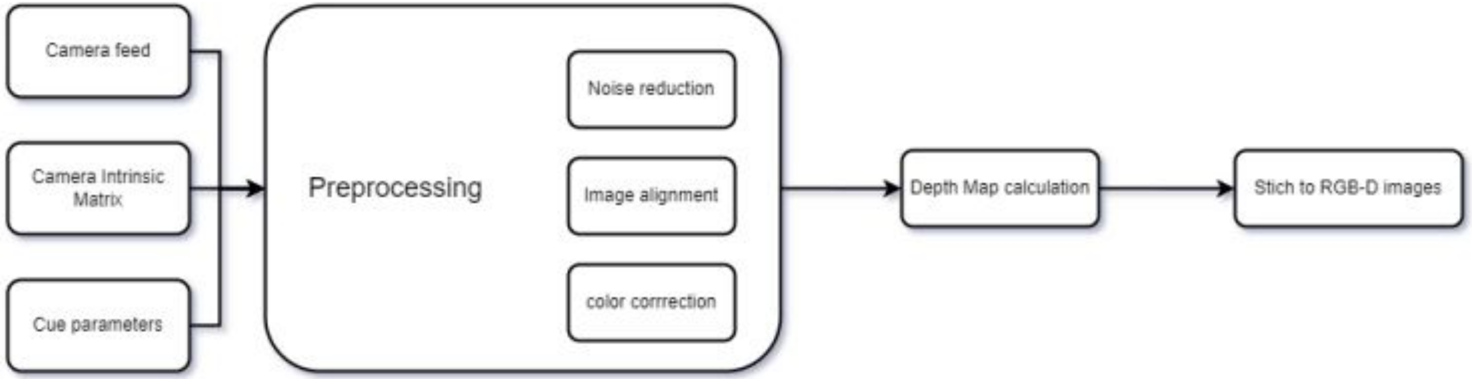


Fig 9: Software block diagram

Methodology

Software Design

- The software for this project tackles the video feed from the camera and turns it into a depth map showing how far away things are.
- First, the program cleans up the video, getting rid of any noise. If multiple pictures are needed, the software align them up perfectly.
- Then the model figures out depth from the video by understanding visual cues.
- Once it has depth information, the program combines it back with the original color picture to create a special image with both color and depth data.

Model Training Pipeline of DPT transformer

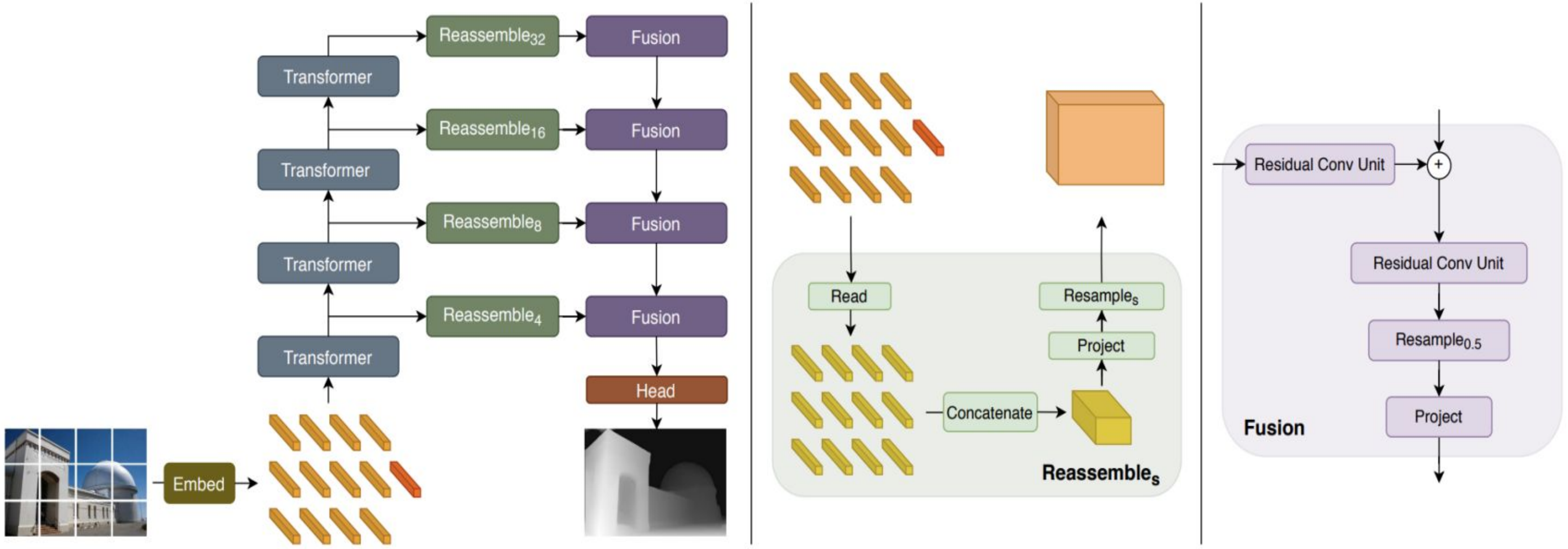


Fig 10 : Supervised learning of DPT transformer

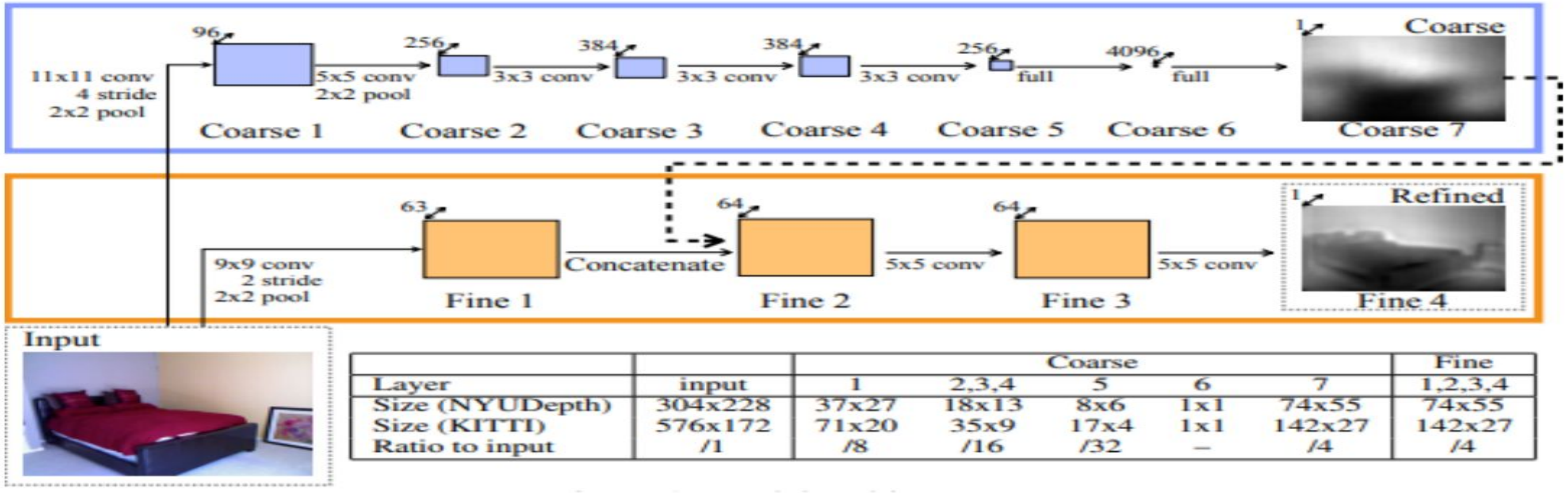
Model Training Pipeline of DPT transformer

- The input image is transformed into tokens using one of two methods:
  - Extracting non-overlapping patches followed by a linear projection of their flattened representation.
  - Applying a ResNet50 feature extractor.
- The image embedding is augmented with a positional embedding.
- A patch-independent readout token is added to the tokens.
- The tokens are passed through multiple transformer stages.
- Tokens from different stages are reassembled into an image-like representation at multiple resolutions.
- Fusion modules progressively fuse and upsample these representations to generate a fine-grained prediction.
- Tokens are assembled into feature maps with 1/8 the spatial resolution of the input image.

Methodology

Model Training Pipeline of Multiscale model.

- The described method involves a global, coarse-scale network with five convolution and max-pooling layers, followed by two fully connected layers for feature extraction.
- Input images are downsampled by a factor of 2, and the final output is at 1/4-resolution of this downsampled input. This output corresponds to a center crop, retaining most of the input image while losing a small border area due to the initial layer of the fine-scale network and image transformations.



		Coarse					Fine
Layer	input	1	2,3,4	5	6	7	1,2,3,4
Size (NYUDepth)	304x228	37x27	18x13	8x6	1x1	74x55	74x55
Size (KITTI)	576x172	71x20	35x9	17x4	1x1	142x27	142x27
Ratio to input	/1	/8	/16	/32	–	/4	/4

Fig 11: Multiscale learning

Methodology

Visual Cues

Stereo Vision (Binocular Disparity)

- Depth can be estimated by calculating the disparity between corresponding points in two images captured from slightly different viewpoints.
- The disparity, or difference in the relative position of an object in the two images, is inversely proportional to the object's distance.

Motion Parallax

- Motion parallax is the optical change in the visual field resulting from a change in the observer's viewing position.
- Nearby objects appear to move faster across the field of view than distant objects, providing depth and distance cues.

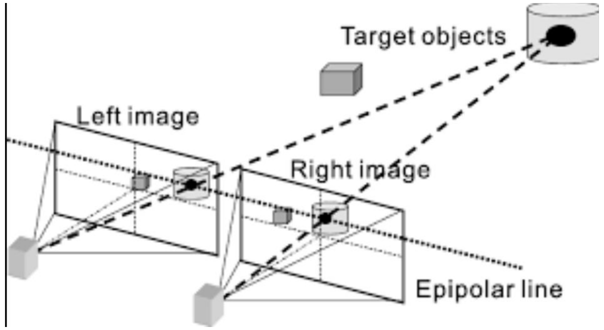


Fig 12: Stereo Vision

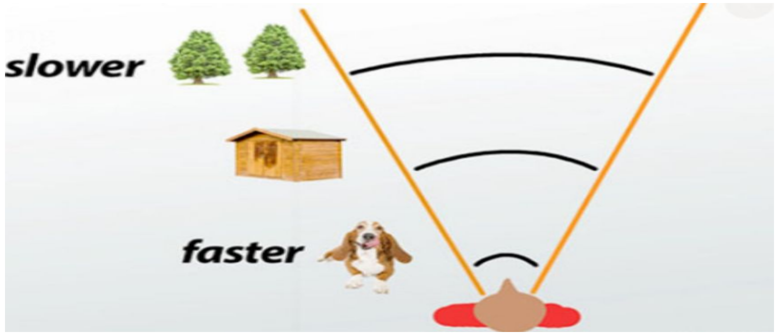


Fig 13: Motion Parallax

Methodology

Visual Cues

Structured Light

- Structured lighting involves projecting a known pattern onto a scene and analyzing the deformation of the pattern from a different angle to determine depth .
- By comparing the projected and observed patterns, a depth map is generated.

Active Illumination

- Active illumination projects light, such as lasers or infrared, onto a scene and measures the reflected light to determine depth .

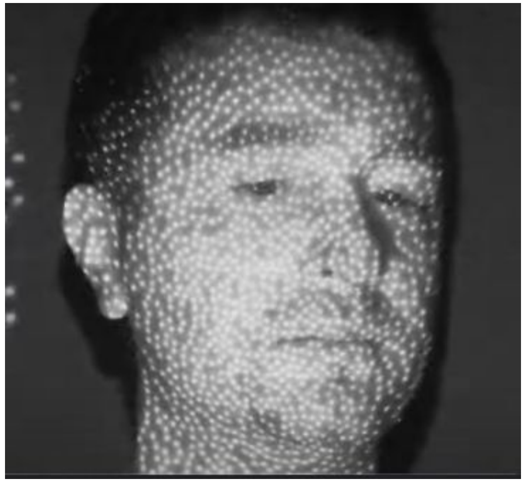


Fig 14: Structured light

Methodology

Other Visual Cues

- Texture Gradient
- Occlusion (Interposition)
- Shading and Shadows
- Aerial Perspective (Atmospheric Perspective)
- Relative Size
- Known Object Size (Size Constancy)
- Linear Perspective
- Defocus Blur (Depth from Defocus)

Time frame of work

The following Gantt chart in figure shows the work flow along with the duration required for the project.

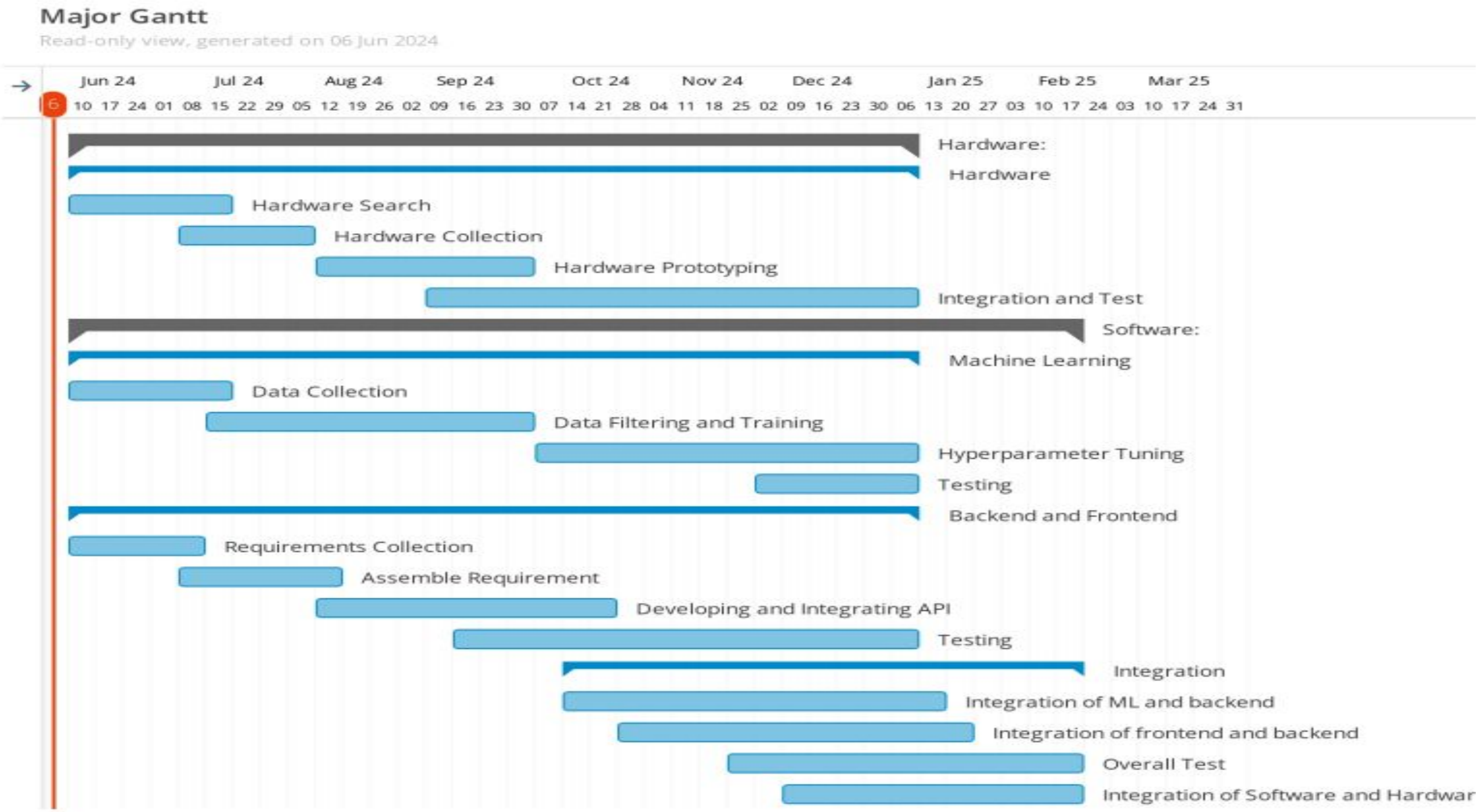
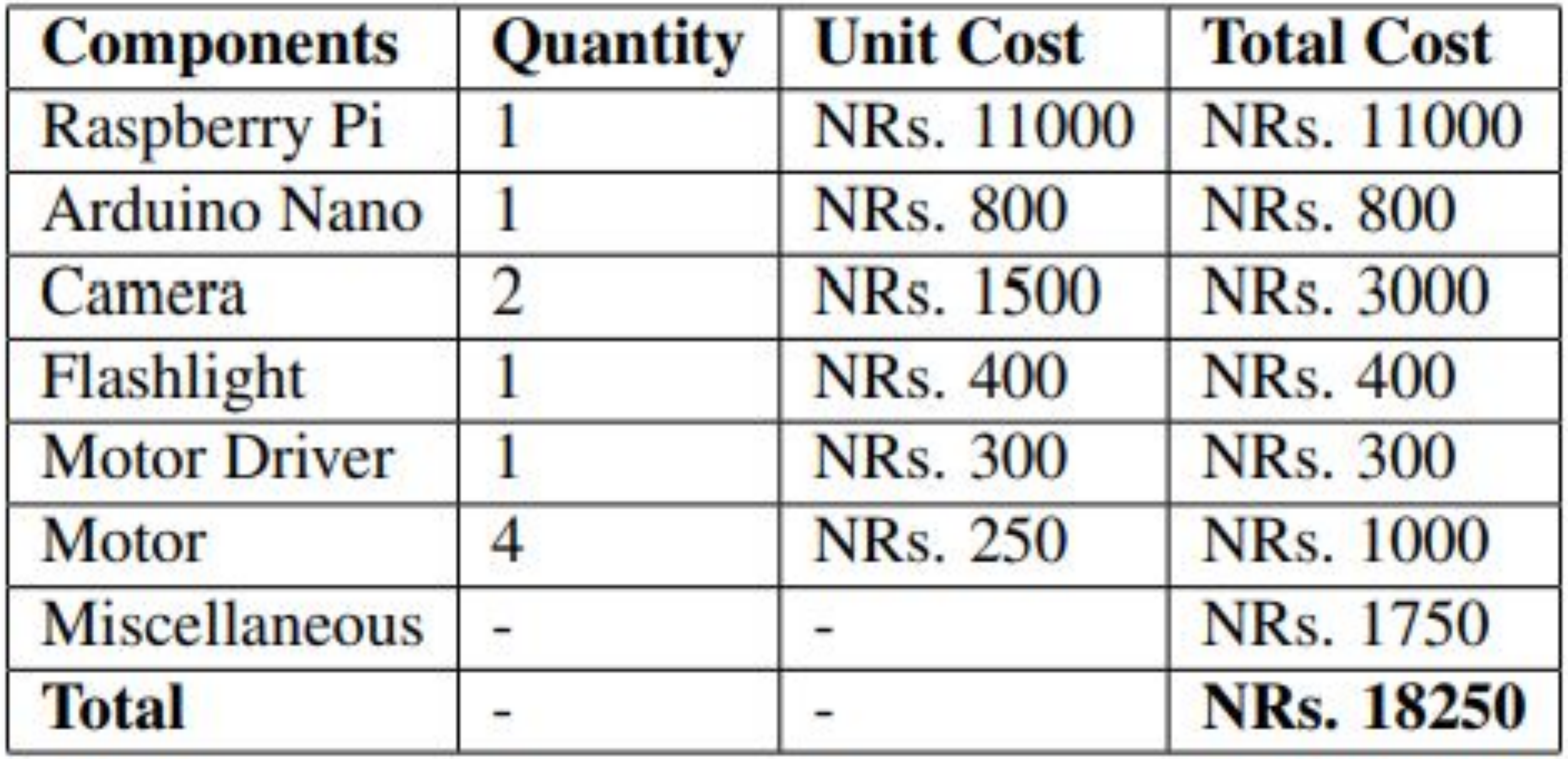


Fig 15: Gantt chart for the time frame of work

Cost Estimation

Cost estimation of the proposed project includes the cost of required materials mentioned in the table .



Components	**Quantity**	**Unit Cost**	**Total Cost**
Raspberry Pi	1	NRs. 11000	NRs. 11000
Arduino Nano	1	NRs. 800	NRs. 800
Camera	2	NRs. 1500	NRs. 3000
Flashlight	1	NRs. 400	NRs. 400
Motor Driver	1	NRs. 300	NRs. 300
Motor	4	NRs. 250	NRs. 1000
Miscellaneous	-	-	NRs. 1750
Total	-	-	**NRs. 18250**

Expected Outcome

The expected output of our system is as follows :

- Generate accurate Depth map from image
- Visualization environment for depth map
- Cost-effective solution of depth compared to LIDAR and TOF sensors

References

[1] M. J. Brooks and B. K. Horn, “Shape and source from shading,” 1985.

[2] D. Marr, “A photogrammetric method for determining the shape of a nonrigid object from a sequence of photographs,” 1975.

[3] K.-i. Kanatani and T.-C. Chou, “Shape from texture: General principle,” Artificial Intelligence, vol. 38, no. 1, pp. 1–48, 1989.

[4] D. Eigen, C. Puhrsch, and R. Fergus, “Depth map prediction from a single image using a multi-scale deep network,” Advances in neural information processing systems, vol. 27, 2014.

[5] G. Liu, G. Jiang, R. Xiong, and Y. Ou, “Binocular depth estimation using convolutional neural network with siamese branches,” in 2019 IEEE International Conference on Robotics and Biomimetics (ROBIO). IEEE, 2019, pp. 1717–1722.

[6] A. Roy and S. Todorovic, “Monocular depth estimation using neural regression forest,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 5506–5514.

[7] C. Godard, O. Mac Aodha, and G. J. Brostow, “Unsupervised monocular depth estimation with left-right consistency,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 270–279.

[8] S. Gasperini, N. Morbitzer, H. Jung, N. Navab, and F. Tombari, “Robust monocular depth estimation under challenging conditions,” in Proceedings of the IEEE/CVF international conference on computer vision, 2023, pp. 8177–8186.

References

[9] H. Fu, M. Gong, C. Wang, K. Batmanghelich, and D. Tao, “Deep ordinal regression network for monocular dept estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 2002–2011. 12

[10] D. Wang, Z. Liu, S. Shao, X. Wu, W. Chen, and Z. Li, “Monocular depth estimation: A survey,” in IECON 2023-49th Annual Conference of the IEEE Industrial Electronics Society. IEEE, 2023, pp. 1–7.

[11] D. Xu, E. Ricci, W. Ouyang, X. Wang, and N. Sebe, “Multi-scale continuous crfs as sequential deep networks for monocular depth estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2017, pp. 5354– 5362.

[12] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe, and E. Ricci, “Structured attention guided convolutional neural fields for monocular depth estimation,” in Proceedings of the IEEE conference on computer vision and pattern recognition, 2018, pp. 3917–3925.

[13] C. Godard, O. Mac Aodha, M. Firman, and G. J. Brostow, “Digging into selfsupervised monocular depth estimation,” in Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 3828–3838. [14] D. Wofk, F. Ma, T.-J. Yang, S. Karaman, and V. Sze, “Fastdepth: Fast monocular depth estimation on embedded systems,” in 2019 International Conference on Robotics and Automation (ICRA). IEEE, 2019, pp. 6101–6108.

References

[15] R. Furukawa, R. Sagawa, and H. Kawasaki, “Depth estimation using structured light flow–analysis of projected pattern flow on an object’s surface,” in Proceedings of the IEEE international conference on computer vision, 2017, pp. 4640– 4648.

[16] “L293x quadruple half-h drivers.” [Online]. Available: https://www.ti.com/product/L293D [17] [Online]. Available: https://flask.palletsprojects.com/en/3.0.x/

[18] W. E. L. Grimson, “A computer implementation of a theory of human stereo vision,” Philosophical Transactions of the Royal Society of London. B, Biological Sciences, vol. 292, no. 1058, pp. 217–253, 1981. 13 [19] E. J. Gibson, J. J. Gibson, O. W. Smith, and H. Flock, “Motion parallax as a determinant of perceived depth.” Journal of experimental psychology, vol. 58, no. 1, p. 40, 1959.

[20] M. Nawrot and K. Stroyan, “The motion/pursuit law for visual depth perception from motion parallax,” Vision research, vol. 49, no. 15, pp. 1969–1978, 2009.

[21] J. Salvi, J. Pages, and J. Batlle, “Pattern codification strategies in structured light systems,” Pattern recognition, vol. 37, no. 4, pp. 827–849, 2004.

References

[22] Z. Cai, X. Liu, G. Pedrini, W. Osten, and X. Peng, “Accurate depth estimation in structured light fields,” Opt. Express, vol. 27, no. 9, pp. 13 532–13 546, 4 2019. [Online]. Available: https://opg.optica.org/oe/abstract.cfm?URI=oe-27-9-13532

[23] R. Ranftl, A. Bochkovskiy, and V. Koltun, “Vision transformers for dense prediction,” in Proceedings of the IEEE/CVF international conference on computer vision, 2021, pp. 12 179–12 188.