Moving Object Detection, Tracking and Range Estimation in Infrared Videos using Deep Learning

Shubham Kasera, Ajay Waghumbare, Sahil Mahajan, Upasna Singh*

School of Computer Engineering & Mathematical Sciences (SoCE&MS), Defence Institute of Advanced Technology, Pune, 411025, Maharashtra, India

* Corresponding author

doi: https://doi.org/10.21467/proceedings.178.28

ABSTRACT

Infrared video technology for moving Object Detection, Tracking and Range Estimation has become a pivotal tool in various fields such as video surveillance, infrared guidance, Unmanned Aerial Vehicle (UAV) based monitoring and autonomous vehicle systems to medical imaging and environmental monitoring. Detecting and estimating the range of a moving object in an infrared video is a critical task with applications in target tracking, obstacle avoidance, and 3D scene reconstruction. This abstract highlight the key aspects of R&D in the field of detecting and estimation of objects tracking and range detection using infrared video. The abstract sheds light on the emerging trends in object range detection, such as Computer Vision (Optical Flow) and Deep Learning for object detection, tracking which is useful for range estimation of a moving object in an infrared images time frames. The utilization of Neural Networks (NN) and Convolutional Neural Networks (CNNs) such as YOLOv8, Mask R-CNN and Faster R-CNN deep learning for Object Detection, LucasKanade and DenseNet Optical Flow techniques for object tracking and MonoDepth deep learning model for Range Estimation in infrared videos along with the proposed model is explored in detail.

Keywords: Infrared videos, Object Detection, Optical Flow, Range Estimation, Computer Vision.

1 Introduction

The study reviews and analyses the methodologies and techniques employed to detect the moving object, track and estimate the range of moving objects in infrared video. This paper discusses the challenges associated with range detection in an infrared video, including environmental factors like ambient temperature, humidity, and atmospheric conditions, which can affect accuracy. In today's rapidly advancing technological landscape, the use of infrared imaging has found applications in a wide range of domains, including surveillance, autonomous vehicles, search and rescue operations, and military reconnaissance. Infrared cameras provide the unique capability to capture images in environments with low or no visible light, making them indispensable for various critical tasks. One of the fundamental challenges in utilizing infrared imagery is accurately estimating the range or distance to objects within the scene. This paper focuses into the fascinating realm of object range estimation in infrared videos. Object range detection involves determining the distance between an infrared camera and an object within its field of view. Achieving accurate range estimation in real-time applications has significant implications for the enhancement of object detection and tracking leading to improved situational awareness and safety in various domains. The primary goal of the study is to explore algorithms and develop the models for improving the accuracy of range estimation in the context of infrared video analysis. In this study, aim is to investigate the impact of moving object range estimation in infrared videos by using various pretrained and proposed Deep Learning models as mentioned in Table 2, Table 4 and Table 5. The study was carried out by using the Deep Learning model such as

Proceedings DOI: 10.21467/proceedings.178; Series: AIJR Proceedings; ISSN: 2582-3922; ISBN: 978-81-984081-8-1

MonoDepth, Faster RCNN etc. and the results were analysed to provide insights into Range Estimation in Infrared Videos.

2 Proposed Methodology

2.1 Moving Object Detection

One of the most significant and difficult subfields in computer vision and deep learning is object detection as discussed in many reports [4], [12], [15], [17] which finds instances of semantic objects of a particular class and is frequently used in applications such as autonomous driving and security monitoring. The performance of object detectors has significantly improved with the quick development of deep learning networks for detection tasks. In this study analysis of the techniques for the most common range detection models are now in use which provide a description of the benchmark datasets in order to fully and clearly grasp the major development state of the object detection pipeline. Next, and above all, this paper gives a thorough, methodical discussion of a range of object detection techniques, focusing on one and two stage detectors. Knowledge transfer is being applied in this paper such as Faster R-CNN [13] and YOLOv8 [8] and the proposed model which will help to verify and predict the object detection results with correct class and the confidence score in infrared dataset video frames.

2.2 Moving Object Tracking

Object tracking, as highlighted in numerous studies [1], [7], [9]-[11] is a key aspect of computer vision, essential for a wide range of applications such as surveillance and augmented reality. The task involves identifying and following one or more objects throughout a sequence of video frames or image streams over time. The primary objective is to accurately predict and track the objects' paths, even in the presence of challenges like occlusions, lighting variations, and background interference. By providing crucial insights into the movement and behaviour of objects, object tracking supports various applications, including video analysis, anomaly detection, and automated monitoring, thereby enhancing the development of intelligent systems. This paper applies the Lucas-Kanade Optical Flow method [7] and the DenseNet deep learning model [6] to assess object tracking performance on infrared video frames.

2.3 Range or Depth Estimation

Range or depth estimation, as explored in several studies [1], [2], [14], involves predicting the distance to objects in a scene from an image or sensor data input. This process is essential for a variety of applications, including autonomous navigation, robotics, augmented reality, and 3D scene reconstruction. Multiple deep learning techniques have been developed for depth estimation, utilizing methods such as Convolutional Neural Networks (CNNs), recurrent neural networks (RNNs), and transformer-based models. The MonoDepth model [5], in particular, employs a neural network to predict the depth of a scene from a single image. This innovative approach uses CNNs trained on large datasets to infer the distance information within the scene. By estimating depth from just one image, the MonoDepth model [3] provides a more efficient solution than traditional stereo methods that require multiple images. The model learns to deduce depth from cues like perspective, texture gradients, and object sizes, allowing it to produce accurate distance estimations. With ongoing advancements in deep learning, the MonoDepth model [6] shows great promise in improving depth perception, which enhances the ability of machines to interact with and understand their environments. The disparity map

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

produced by the MonoDepth model visually represents the differences in pixel locations between corresponding points in stereo image pairs. In stereo vision, two cameras capture the same scene from slightly different angles, similar to how human eyes perceive depth. By comparing pixel intensities from both images, a stereo vision system calculates the horizontal displacement (disparity) of each point between the two views. This disparity map is then used to determine the depth of objects in the scene. The architecture of the MonoDepth2 model [9] is illustrated in Fig. 1.

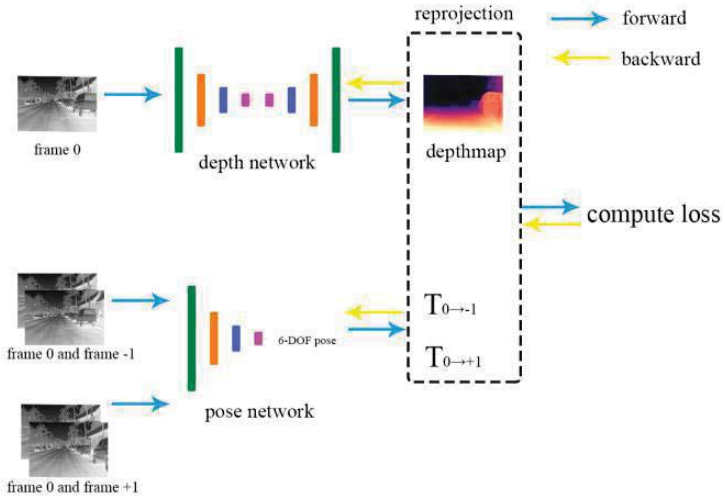


Figure 1: Architecture for Range (Depth) Estimation model

2.4 Integration of Moving Object Detection, Object Tracking and Range Estimation

As discussed in the report [4], object detection is the task of identifying and locating objects within an image or video frame. This process involves algorithms that classify objects and outline them with bounding boxes. Once the objects are detected, the next step is object tracking, as described in [9], which involves following the identified objects across consecutive frames while maintaining their identities over time. This is essential for understanding how objects move and behave in dynamic environments. Range estimation, as explained in [14], refers to determining the distance of objects from the observer or camera, providing important spatial data that enhances the understanding of object positioning and their spatial relationships in a 3D scene. These three components object detection, tracking, and range estimation can be integrated to improve overall scene understanding through several processes such as initialization, association, contextual information, and feedback loops. Object detection can be used to initiate object tracking by identifying objects of interest in the scene. Similarly, range estimation can offer initial depth data to help initialize tracking algorithms. Object tracking then uses the output of detection to associate tracked objects with newly detected ones in subsequent frames, ensuring continuous object identification. Range estimation adds valuable contextual data about the scene's geometry, which can enhance both detection and tracking processes. For instance, depth information helps validate the size and location of detected objects, improving the precision of the detection process. It also aids tracking by offering predictions for object movement based on depth cues. Furthermore, the results from object detection, tracking, and range estimation can be integrated into a feedback loop, refining each component's outcomes iteratively. For example, motion data from tracking can support more accurate range estimation, while depth information from range estimation can improve the tracking predictions.

3 Experimental Setup

3.1 Implementation Platform

The overall deep learning model for Object Detection, Tracking and Range Estimation is trained, compiled and implemented on Google Colab platform by utilising the T4 GPU. The TensorFlow version is 2.15.0 and the Python version is 3.10.12.

3.2 Infrared Dataset

In this study, training, validation and testing of the suggested strategy have been accomplished by using the Dimensional small target detection and tracking infrared dataset as mentioned in [16] . The video infrared datasets is clipped into different frames which helps in maintaining the accuracy and for the better performance of the model. The dataset consists of total of 6 classes and almost 5000 image time frames. Some of the dataset samples are shown in the below Table 1. The Infrared dataset classes wise frames count are shown in Fig.2.

Table 1: *IR Dataset samples and its classes*

(a) Sky	(b) Forest	(c) Mid Air
		
(d) Hills	**(e) Mountain**	**(f) Plain**
		

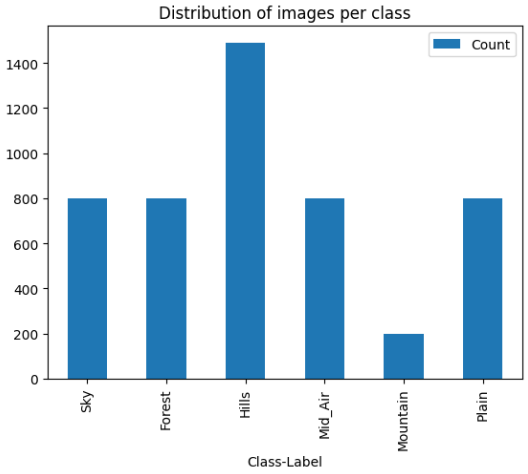


Figure 2: Infrared dataset classes wise frames count

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

4 Experiment and Results

4.1 Moving Object Detection

In this paper the three Deep Learning architectures are under investigation, i.e. Faster R-CNN model [14], YoloV8 model [12] and the Proposed Model. The Faster R-CNN and YoloV8 are the pre trained models used to verify the object detection result with the proposed model which is implemented and discussed in this paper. The number of objects that can be detected are UAV, aeroplane, aircraft, tree, person and bird.

4.1.1 Faster R-CNN

Faster R-CNN as mentioned in the [14] (Region-based Convolutional Neural Network) is a deep learning model developed for object detection tasks. Faster R-CNN improves upon its predecessor, R-CNN, by integrating the region proposal network (RPN) directly into the model. This eliminates the need for external region proposal methods, making the detection process faster and more efficient. Faster R-CNN achieves state-of-the-art performance in terms of both speed and accuracy in object detection tasks, making it widely used in various applications such as autonomous driving, surveillance, and medical imaging. The key components of Faster R-CNN include:

a) **Region Proposal Network (RPN):** A Region Proposal Network (RPN) is a fully convolutional network used to generate potential object locations (bounding boxes) within an image. This network is a crucial component in many object detection models, such as Faster R-CNN. The RPN scans the input image using a small network to produce a set of bounding box proposals, each representing a region that might contain an object. These proposed regions are then passed on to a second network for further classification and refinement. In addition to providing class scores for each region, the RPN also predicts the necessary adjustments to the bounding box to better match the object's location. The RPN significantly improves the efficiency of object detection by eliminating the need for traditional, time-consuming search methods like selective search.

b) **Region of Interest (ROI) Pooling:** After generating the region proposals, the Region of Interest (ROI) extract the features from each region and adjusts to a consistent size for further analysis.

c) **Object Detection Network:** The object detection network leverages the ROI features to determine class labels and fine-tune the bounding box coordinates for each region proposal.

4.1.2 YOLOv8

YOLOv8 [12] is a leading deep learning model designed for real-time object detection in computer vision applications. Its sophisticated architecture and cutting-edge algorithms have transformed object detection, enabling highly accurate and efficient detection of objects in real-time settings.

4.1.3 Proposed Object Detection Model

The proposed moving Object Detection model implemented in this paper also known as Mask R-CNN deep learning model as mentioned in paper [17] as a part of Detectron framework. Detectron is a stateof-the-art object detection and segmentation framework developed by Facebook AI Research (FAIR). It provides a flexible and efficient platform for training and deploying deep learning models for various computer vision tasks, including object detection, instance segmentation, and keypoint detection. The Infrared dataset is completely trained on the model before performing the object detection. The proposed deep learning model architecture diagram is shown in Fig.3 and the model training accuracy and loss curves are given in Fig.4 and Fig.5.

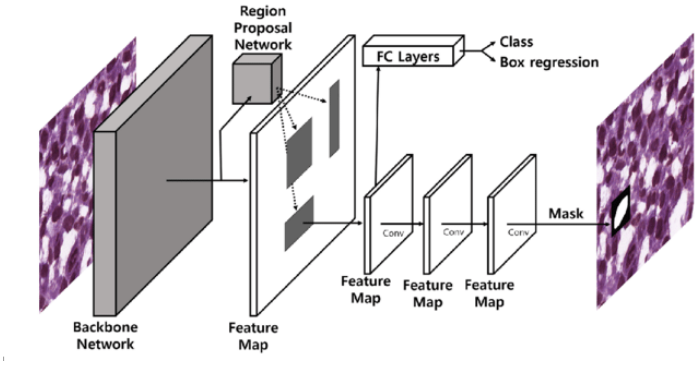


Figure 3: Proposed Object Detection model architecture diagram

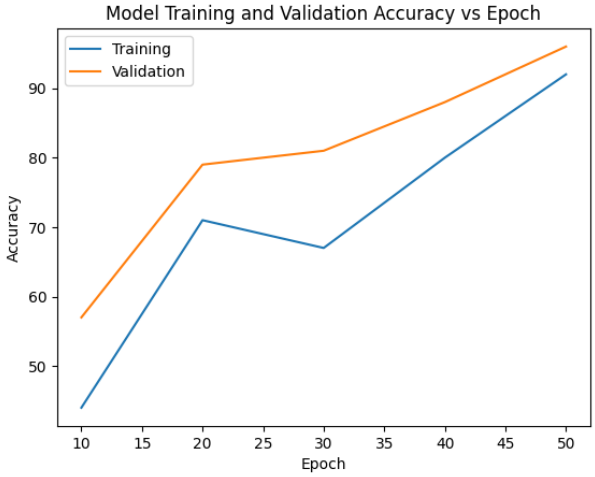


Figure 4: Model Training and Validation Accuracy vs Epoch graph

Figure 5: Model Training and Validation Loss vs Epoch graph

The results for object detection using the pre trained model as mentioned and the proposed model are discussed in Table 2. The result evaluation of moving Object Detection w.r.t the pre trained deep learning models Faster R-CNN & YOLOv8 and the proposed model are discussed briefly in Table 3. As per Table 3, the results of the proposed model are more than the pre trained models as mentioned for Faster R-CNN and YOLOv8 deep learning models. Hence in this study the results for Object

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

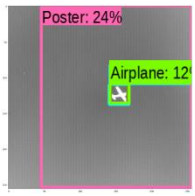
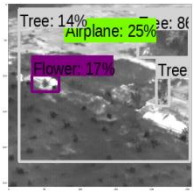
Series: AIJR Proceedings
ISSN: 2582-3922

Tracking and Range Estimation can be considered for Object Detection. The tabulated results using the proposed model will help in the further calculation and estimation of range in infrared datasets.

4.2 Object Tracking

Object tracking in infrared datasets is a crucial element in computer vision and deep learning, enabling the observation and analysis of moving objects over time. Two popular techniques for object tracking are the Lucas-Kanade Optical Flow method and the Dense Optical Flow method. The Lucas-Kanade Optical Flow method is a well-established and efficient algorithm used for optical flow estimation. It works by examining small windows around each pixel, calculating flow vectors that describe the apparent motion of those windows between consecutive frames. By solving a set of linear equations, it locally estimates motion parameters, making it effective for tracking objects with smooth movements. However, it may face difficulties with large displacements or complex motion patterns. On the other hand, the DenseNet Optical Flow method calculates optical flow vectors for every pixel in an image, providing a comprehensive analysis of motion across the entire frame. This method offers more detailed motion information than the Lucas-Kanade approach and can handle intricate motion patterns and occlusions. Dense optical flow estimates motion vectors for 4x4 pixel blocks between consecutive frames. Although this method delivers more precise motion analysis, it demands higher computational resources due to the dense nature of the calculations. In this paper, the Densely Connected Convolutional Networks (DenseNet) model is applied alongside deep learning techniques. DenseNet is a type of Convolutional Neural Network (CNN) architecture in which each layer is connected to every other layer in a feed-forward manner, allowing later layers to leverage features from earlier ones. DenseNet consists of dense blocks and transition layers, where each layer within a dense block is interconnected with all other layers. The results for object tracking are provided in Table 4 below.

Table 2: *Moving Object Detection results on IR datasets using Faster R-CNN, YOLO-v8 Pretrained models and the proposed model*

Sl. No.	Dataset Classes	Faster R-CNN	YoloV8	Proposed Model
1.	Sky			
2.	Forest			
3.	Mid Air			

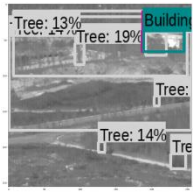
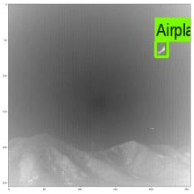
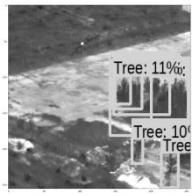
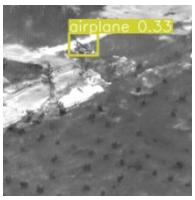
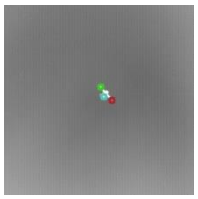
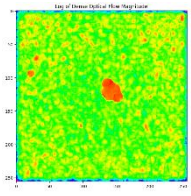
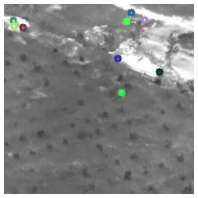
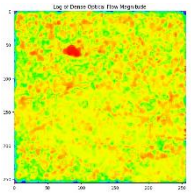
4.	Hills			
5.	Mountain			
6.	Plain			

Table 3: *Result Evaluation of Object Detection Confidence Scores obtained on IR dataset*

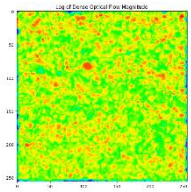
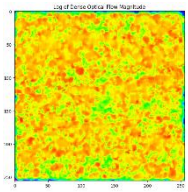
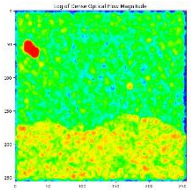
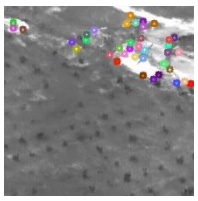
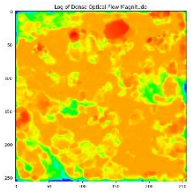
Sl. No.	Dataset Classes	Frames	Faster R-CNN	YoloV8	Proposed Model
1.	Sky	798	0.12	0.7	0.97
2.	Forest	798	0.25	0.37	0.44
3.	Mid Air	1490	0.1	0.26	0.69
4.	Hills	798	0.19	0.18	0.16
5.	Mountain	200	0.29	0.26	0.81
6.	Plain	798	0.36	0.33	0.63

Table 4: *Moving Object Tracking results on IR datasets using Lucas-Kanade Optical Flow and DenseNet Deep learning based method*

Sl. No.	Dataset Classes	Lucas-Kanade Optical Flow Method	DenseNet Method
1.	Sky		
2.	Forest		

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

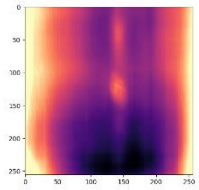
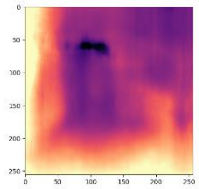
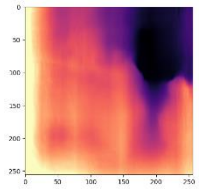
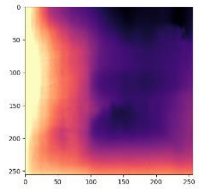
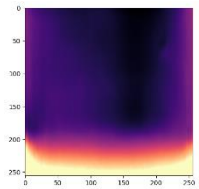
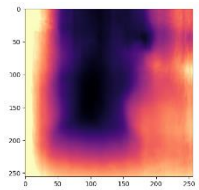
Series: AIJR Proceedings
ISSN: 2582-3922

3.	Mid Air		
4.	Hills		
5.	Mountain		
6.	Plain		

4.3 Range Estimation

Range or depth estimation using deep learning techniques has emerged as a pivotal area within computer vision, offering solutions with widespread applications in fields such as robotics, autonomous driving, and augmented reality. Encoders have become popular choices for extracting hierarchical features from images, while decoders reconstruct depth maps or distance estimations from these features. Despite significant advancements, challenges such as handling occlusions, varying lighting conditions, and generalization to unseen environments persist. Nonetheless, ongoing R&D endeavours focus on novel architectures, loss functions and data augmentation techniques to further enhance the accuracy and robustness of range or depth estimation systems, promising continued progress in this dynamic field. In this paper a deep learning model proposed is known as MonoDepth for the Range Estimation. It is a deep learning model designed for monocular (single camera) depth estimation, a crucial task in computer vision with applications in various fields like robotics, augmented reality, and autonomous driving. This model is built upon the principles of self-supervised learning, leveraging large-scale datasets to train the network without requiring explicit depth annotations. MonoDepth effectively learns to infer depth information, capturing intricate details of the scene's structure and depth variations. The model has almost 14.84 million parameters which generates the disparity map which help to find the depth of the particular pixel value in meters. The results for Range / Depth estimation with the actual and predicted depth values in meters along with the absolute error are mentioned in Table 5.

Table 5: *Depth Estimation Results and Error evaluation in meters (m)*

Sl. No.	Dataset Classes	Depth Map	Object Location (x, y)	Actual Depth (m)	Predicted Depth (m)	Absolute Error (m)
1.	Sky		(160,124)	154	145	9
			(148,130)	147	139	8
2.	Forest		(205,74)	174	168	6
			(190,81)	158	149	9
3.	Mid Air		(212,176)	114	103	11
			(207,178)	110	101	9
4.	Hills		(216,166)	159	151	8
			(208,167)	157	147	10
5.	Mountain		(399,139)	264	255	9
			(386,512)	261	253	8
6.	Plain		(367,185)	163	157	6
			(375,181)	185	178	7
					Average Error (m)	**8.34**

4.4 Range (Depth) Estimation Formula and Equations

The Range and the depth estimation formula are mentioned below as the relative distance shown in Equation no. 1 Depth information shown in Equation no. 2 and the depth value calculation shown in Equation no. 3 formula are shown mathematically below.

$$Relative\ Distance\ (meters) = \frac{(focal\ length\ x\ baseline)}{Depth\ Value} \quad (1)$$

The focal length is the infrared camera focal length in pixels. Base line is the distance between left and right cameras in the case of stereo (dual) cameras. In this paper Monocular (single) based infrared cameras were used for which the baseline constant value is 1. And the depth value is nothing but the

Proceedings DOI: 10.21467/proceedings.178
ISBN: 978-81-984081-8-1

Series: AIJR Proceedings
ISSN: 2582-3922

particular pixel or depth value in meters of a particular coordinates within the depth map image generated by the depth estimation or range estimation deep learning based model.

$$Depth\ Information = \frac{1}{Depth\ map\ pixel\ value} \quad (2)$$

The pixel value is the pixel depth value of the particular coordinates within the depth map generated by the proposed deep learning model.

$$Depth\ Value\ (meters) = \frac{1}{(Depth\ Image\ value\ x\ Scaling\ Factor)} \quad (3)$$

The Depth Image Value is the depth value of a particular coordinate within the depth map hence generated. The scaling factor is the constant value that acts as a baseline for the estimation of range or depth using the disparity or the depth map. In this study Monocular based infrared based thermal cameras were used for which the scaling factor is 1. As mentioned in Table 5, the depth map color bar represents the distance or depth of objects in a scene from the viewpoint of a camera or sensor. In a depth map, closer objects are usually depicted in warmer colors (such as red or yellow), while farther objects are represented in cooler colors (such as blue or purple). The color magma gradient representation ass shown in Table 5, helps to visually differentiate between objects at varying distances, making it easier to interpret the spatial relationships within the scenes of infrared dataset.

5 Conclusions

This paper contributes towards the understanding, analysis and enhancement of moving Object Detection, Tracking and Range Estimation in an Infrared video dataset. The Faster R-CNN model has been fine tuned for the moving Object Detection and the same is proposed to get the better accuracy results. Out of all the 06 infrared dataset classes the Sky class gives the maximum result of 97%. The DenseNet model shows better results than the Lucas-Kanade Method for the Object Tracking. The moving object has been tracked and indicated with the darker color regions as shown in Table 4. A model is proposed for range estimation to generate the depth or disparity map. Then the disparity map is loaded onto the GUI tool implemented and developed to get the depth of a particular pixel of an object within the image as shown in table 5. In summary by integrating Object Detection, Object Tracking and Range Estimation, Computer Vision and Deep Learning systems can achieve robust scene understanding, enabling applications that require accurate perception and interaction with the environment. The depth estimation model implemented can be considered and deployed in the real time applications such as Surveillance and Security, Robotics, Augmented Reality and Autonomous Vehicles.

6 Declarations

6.1 Competing Interests

The author declares that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

6.2 Publisher's Note

AIJR remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

How to Cite

Shubham Kasera, Ajay Waghumbare, Sahil Mahajan, Upasna Singh (2025). Moving Object Detection, Tracking and Range Estimation in Infrared Videos using Deep Learning. *AIJR Proceedings*, 263-274. https://doi.org/10.21467/proceedings.178.28

References

1. Adz-Dzikri, A. A., Virgono, A., & Dirgantara, F. M. (2021, October). Advance Driving Assistance Systems: Object Detection and Distance Estimation Using Deep Learning. In 2021 8th International Conference on Electrical Engineering, Computer Science and Informatics (EECSI) (pp. 381-386). IEEE.
2. Godard, C., Mac Aodha, O., & Brostow, G. J. (2017). Unsupervised monocular depth estimation with left-right consistency. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 270-279).
3. Shin, U., Park, J., & Kweon, I. S. (2023). Deep depth estimation from thermal image. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 1043-1053).
4. Chen, Z., Khemmar, R., Decoux, B., Atahouet, A., & Ertaud, J. Y. (2019, July). Real time object detection, tracking, and distance and motion estimation based on deep learning: Application to smart mobility. In 2019 Eighth International Conference on Emerging Security Technologies (EST) (pp. 1-6). IEEE.
5. Sun, S., Li, L., & Xi, L. (2012, December). Depth estimation from monocular infrared images based on bp neural network model. In 2012 International Conference on Computer Vision in Remote Sensing (pp. 237-241). IEEE.
6. Dosovitskiy, A., Fischer, P., Ilg, E., Hausser, P., Hazirbas, C., Golkov, V., ... & Brox, T. (2015). Flownet: Learning optical flow with convolutional networks. In Proceedings of the IEEE international conference on computer vision (pp. 2758-2766).
7. Shah, S. T. H., Xuezhi, X., & Ahmed, W. (2021). Optical flow estimation with convolutional neural nets. Pattern Recognition and Image Analysis, 31, 656-670.
8. Redmon, J., Divvala, S., Girshick, R., & Farhadi, A. (2016). You only look once: Unified, real-time object detection. In Proceedings of the IEEE conference on computer vision and pattern recognition (pp. 779-788).
9. Parhizkar, M., Karamali, G., & Abedi Ravan, B. (2023). Object tracking in infrared images using a deep learning model and a target-attention mechanism. Complex & Intelligent Systems, 9(2), 1495-1506.
10. Kwan, C., & Budavari, B. (2020). Enhancing small moving target detection performance in low-quality and longrange infrared videos using optical flow techniques. Remote Sensing, 12(24), 4024.
11. Kale, K., Pawar, S., & Dhulekar, P. (2015, September). Moving object tracking using optical flow and motion vector estimation. In 2015 4th international conference on reliability, infocom technologies and optimization (ICRITO)(trends and future directions) (pp. 1-6). IEEE.
12. Kwan, C., & Larkin, J. (2021, September). Detection of small moving objects in long range infrared videos from a change detection perspective. In Photonics (Vol. 8, No. 9, p. 394). MDPI.
13. Jiang, C., Ren, H., Ye, X., Zhu, J., Zeng, H., Nan, Y., ... & Huo, H. (2022). Object detection from UAV thermal infrared images and videos using YOLO models. International Journal of Applied Earth Observation and Geoinformation, 112, 102912.
14. Rapanotti, J. L., & DEFENCE RESEARCH AND DEVELOPMENT CANADA VALCARTIER (QUEBEC). (2007). Vehicle DAS considerations for the Iron Gorget threats. Analysis, modelling and comments to operations researchers.
15. Aradhya, H. R. (2019). Object detection and tracking using deep learning and artificial intelligence for video surveillance applications. international journal of advanced computer science and applications, 10(12). 16. Bingwei Hui, Zhiyong Song, Hongqi Fan, et al. (2019) . A dataset for infrared image dim-small aircraft target detection and tracking under ground / air background. V1. Science Data Bank. https://doi.org/10.11922/sciencedb.902.
17. He, K., Gkioxari, G., Dollar, P., & Girshick, R. (2017). Mask r-cnn. In Proceedings of the IEEE international confer- ´ ence on computer vision (pp. 2961-2969).