Optimizing Object Classification in Robotic Perception Environments Exploring Late Fusion Strategies

[1]Rodney Adam and [2]Anandakumar Haldorai
[1] School of computer science, The University of Sydney, Camperdown NSW 2050, Australia
[2] Department of Computer Science and Engineering, Sri Eshwar College of Engineering, Coimbatore, India.
[1]adamscse2050@gmail.com, [2] anandakumar.psgtech@gmail.com

Correspondence should be addressed to Rodney Adam**:** adamscse2050@gmail.com

Article Info
Journal of Robotics Spectrum (https://anapub.co.ke/journals/jrs/jrs.html)
Doi: https://doi.org/10.53759/9852/JRS202402008
Received 02 February 2024; Revised from 28 March 2024; Accepted 25 May 2024.
Available online 02 June 2024.

Abstract – Robotic perception systems often include approaches that can extract valuable features or information from the studied dataset. These methods often involve the application of deep learning approaches, such as convolutional neural networks (CNNs), for processing of images, as well as the incorporation of 3D data. The notion of image categorization is well delineated via the use of networks that include convolutional networks. However, some network topologies exhibit a substantial scope and need significant amounts of time and memory resources. On the other hand, the neural networks FlowNet3D and PointFlowNet have the capability to accurately predict scene flow. Specifically, these networks are capable of estimating the three-dimensional movements of point clouds (PCs) within a dynamic environment. When using PCs in robotic applications, it is crucial to examine the robustness of accurately recognizing the points that belong to the object. This article examines the use of robotic perception systems inside autonomous vehicles and the inherent difficulties linked to the analysis and processing of information obtained from diverse sensors. The researchers put out a late fusion methodology that integrates the results of many classifiers in order to enhance the accuracy of categorization. Additionally, the authors propose a weighted fusion technique that incorporates the proximity to objects as a significant factor. The findings indicate that the fusion methods described in this study exhibit superior performance compared to both single modality classification and classic fusion strategies.

Keywords – Object Classification, Intelligent Robotic Perception System, Robotic Perception Environments, Late Fusion Strategies, Deep Learning, Convolutional Neural Networks.

I. INTRODUCTION

In the field of robotics, perception is a comprehensive system that grants the robot the capacity to see, interpret, and engage in reasoning processes pertaining to its immediate surroundings [1]. The fundamental elements of a perception system consist of sensory data processing, data representation (also known as environment modeling), and machine learning-based algorithms, as seen in **Fig 1**. Provided the present state of real-world robotics applications, the focus of this chapter is on weak AI, namely typical machine learning algorithms.

The ability of a robot to perceive its surroundings is of utmost importance in enabling it to make informed choices, strategize, and effectively navigate real-world settings. This is achieved via a wide range of functions and activities, including but not limited to occupancy grid mapping and object identification. Several subareas of robotic perception, such as obstacle detection, semantic place classification, object recognition, voice and gesture recognition, activity classification, road detection, terrain classification, vehicle detection, object tracking, pedestrian detection, environment change detection, and human detection, can be observed in autonomous robot-vehicles. In contemporary times, the majority of robotic vision systems use machine learning (ML) methodologies, including both traditional and deep-learning methodologies. Machine learning techniques are used in robotic perception, taking the form of many approaches such as supervised classifiers, unsupervised learning using manually designed features, deep-learning neural networks such as convolutional neural networks (CNN), or a hybrid mix of numerous methodologies.

These components include sensory data processing, with a particular emphasis on visual and range perception. Additionally, the system incorporates data representations that are tailored to the specific tasks being performed. Furthermore, the system utilizes algorithms that employ artificial intelligence and machine learning techniques for data

analysis and interpretation. Lastly, the system encompasses the planning and execution of actions necessary for facilitating interaction between the robot and its environment. The functionality of robot perception, such as localization and navigation, is contingent upon the specific environment in which the robot is deployed [2]. In essence, a robot is specifically engineered to function within two distinct classifications of habitats, namely indoor and outdoor settings. Hence, it is possible to integrate various assumptions into the mapping and perception systems while considering outdoor or interior situations. Furthermore, the sensors selection is contingent upon the specific habitat, resulting in variations in the sensory input that must be processed by a perception system in indoor and outdoor circumstances.

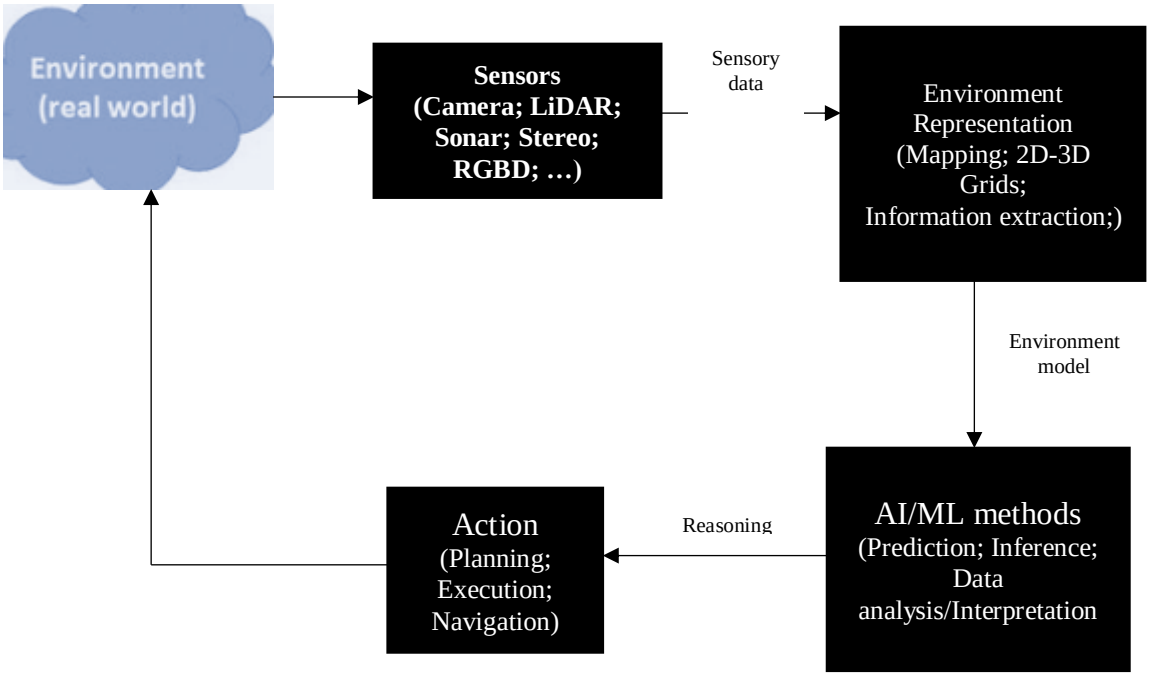


Fig 1. The Essential Components of a Standard Robotic Perception System.

An illustrative instance for elucidating the disparities and complexities of a mobile robot traversing an interior vs outdoor setting lies in the ground, or topography, on which the robot moves. The majority of indoor robots operate under the assumption that the ground is uniform and level, which simplifies the process of creating models to represent the environment. However, outdoor robots, designed for field use, frequently encounter terrain that is irregular and uneven. Consequently, modeling the environment becomes a challenging task, and without an accurate representation, subsequent perception tasks are adversely impacted. Additionally, in outside environments, robotic perception must contend with the challenges posed by weather conditions as well as fluctuations in light intensities and spectrum.

The concept of multi-sensor perception pertains to the capacity of a system or device to collect data from several sensors and amalgamate the information in order to provide a more extensive and precise comprehension of the surrounding world. In a multitude of domains, like autonomous, robotics, and surveillance systems, it is common practice to utilize multiple sensors for the purpose of capturing diverse forms of data. These sensors encompass visual information obtained from cameras, depth information acquired through LiDAR (Light Detection and Ranging) [3], and positional data derived from GPS (Global Positioning System) [4]. The integration of input from many sensors enables a system to boost its perceptual capacities and improve its capacity to perceive and navigate its environment.

The use of an integrated strategy facilitates enhanced decision-making and heightened dependability by mitigating the constraints or uncertainties that might result from depending only on a single sensor type. The integration of many sensors plays a pivotal role in performing activities like as object identification, navigation, and situational awareness across a range of technological applications. The ability of robots to perceive, interpret, and reason about their environment is of utmost importance in the field of robotic perception. The process includes the processing of sensory input, the representation of data, and the use of machine learning algorithms. The integration of input from several sensors, known as multi-sensor perception, serves to augment the system's comprehension of the surrounding world. Late fusion approaches, which include the integration of the outputs from many classifiers, have shown potential in enhancing object categorization inside robotic perception systems.

The primary objective of this article is to provide a valuable contribution to the progress of late fusion techniques in the field of object classification. This research focuses on the incorporation of object distance as a significant weighting component. The rest of the article has been arranged as follows: Section II presents a review of previous literature works

related to the concepts in this research. Section III presents a schematic and systematic methodology employed in composing this paper. Section IV discusses the advancements in robotic perception systems: deep learning, transfer learning, and fusion techniques. Section V presents a detailed analysis of the results, which focus on objects and dataset distance distribution, weighted object distance fusion, and late fusion techniques. Section VI draws a conclusion to the paper, and proposes future research directions.

II. LITERATURE REVIEW

According to Sun, Zhao, and Ma [5], robots have emerged as a significant component due to their capacity to supplant human involvement in both rudimentary and hazardous tasks. Vision-guided robots are often used in several sectors due to their ability to effectively navigate in diverse environments by using input acquired from vision sensors. These robotic systems enhance production by exhibiting remarkable adaptability and robustness. A vision-guided system encompasses many modules, including perception, localization, route planning, and control. In the realm of robot planning, the capacity to effectively respond to abrupt transformations in the environment or navigate around barriers is a significant challenge. Unlike humans, who possess inherent capabilities to effortlessly do these tasks. In the field of robot planning, the process of devising a series of activities from an initial point to a desired target point is facilitated using planning algorithms.

These algorithms serve the purpose of concurrently circumventing any encountered impediments, as discussed by Zeng, Zhang, Chen, Chen, and Liu [6]. The paramount concern in the development of intelligent robots is in the formulation of a highly effective navigation system. Hence, the utilization of vision sensors in the context of robot planning presents a very captivating and expansive field of study. The overarching objective is to attain a secure and optimum path for navigation of the robot. Vision-based robotic systems have numerous applications in various industries, such as spray painting, place and pick operations, assembly tasks in the optical firm, automotive manufacturing, including spot and pipe welding. These systems are also utilized for payload identification and other tasks that require efficient planning and control. The diagram in **Fig 2** depicts the many modules included by vision-based autonomous robotic systems.

According to Macias-Garcia, Galeana-Pérez, Medrano-Hermosillo, and Bayro–Corrochano [7], the perception stage in robotics encompasses the processes of seeing and responding to changes in the immediate surroundings. This is achieved by using the characteristics acquired which aids in decision-making, and execution within real-world settings. The fundamental components of a robotic system of perception primarily consist of the gathering of sensor data and the pre-processing of this data using image processing techniques, enabling the generation of an environment model. Perception refers to the cognitive process of using sensor readings to generate inferences on the surrounding environment.

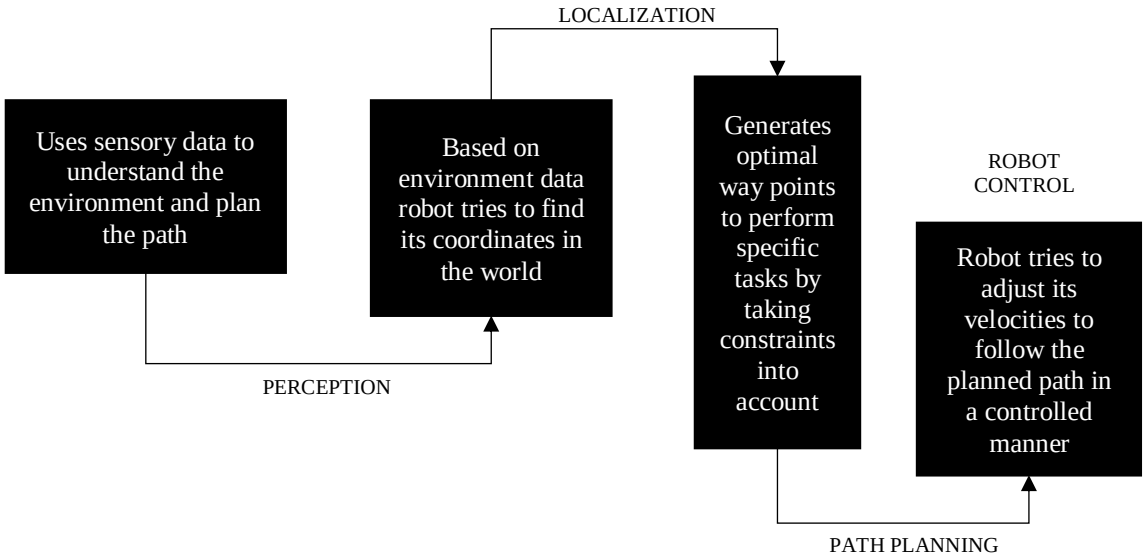


Fig 2. Diagram of the Vision-Based Autonomous Robots' Process Flow.

According to Dong [8], sensors possess observation models that establish a correspondence between the state of the world and the values they measure. These models are sometimes referred to as forward models. In the context of a certain state of the world, the principles of physics may be used to ascertain the sensor output with a high degree of accuracy, accounting for any inherent uncertainties. The concept of perception revolves on the inverse problem, which involves deducing information about the state of the world based on a given set of sensor readings. Inverse issues are well recognized for their inherent difficulty, mostly due to their frequent lack of precise definition.

As stipulated by According to Buosciolo, Pesce, and Sasso [9], in the scenario when noise is absent, it may be inferred that if a range sensor is placed at a 10 meters distance from a wall, the anticipated sensor value would be denoted as z=10.

The objective of the inverse problem is to ascertain certain characteristics or properties of the state of the world, based on the sensor measurement which has been determined to be z=10. The issue at hand lacks clear definition, since it encompasses several potential states of the world that may result in a measurement of z=10. The robot's distance from a wall might potentially be 10 meters, or alternatively, it is plausible that another robot has traversed a distance of 10 meters in close proximity to the sensor. Additionally, it is conceivable that an open door situated 10 meters ahead of the robot has just been closed. The field of perception is primarily focused on addressing these inverse difficulties.

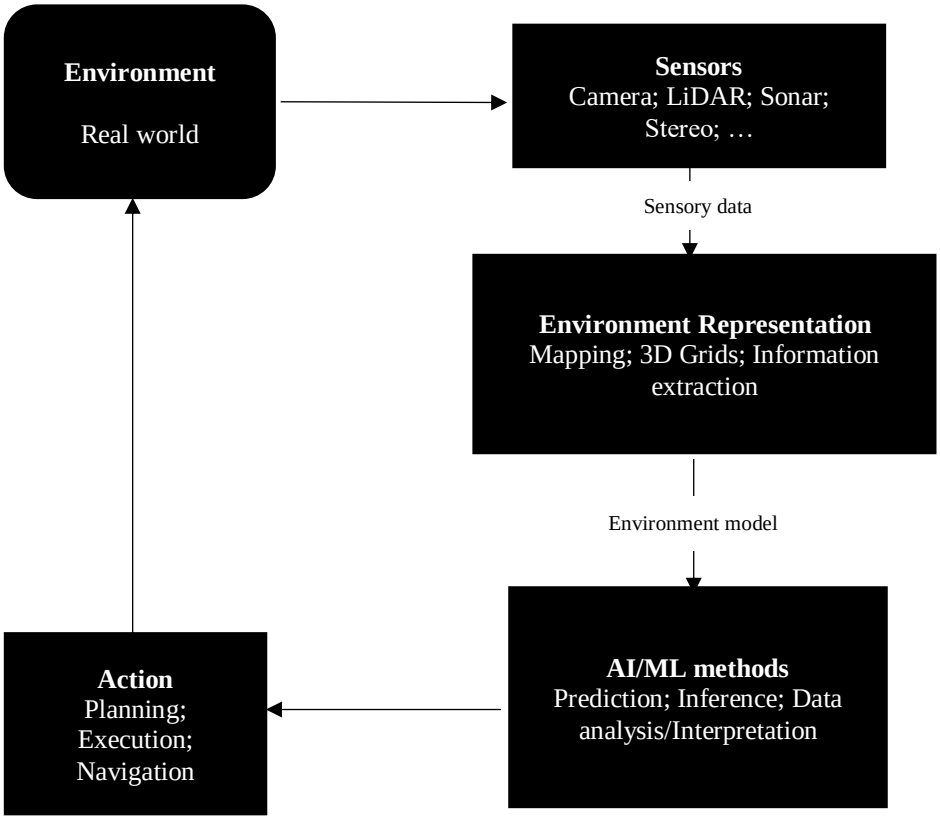


Fig 3. Intelligent Robotic Perception System Components

According to Nesnas, Fesq, and Volpe [10], there has been a notable advancement in the intelligence and autonomy of robots. However, it is important to acknowledge that despite these advancements, robots are still incapable of engaging in cognitive processes, expressing emotions, and exercising human-like decision-making abilities. They lack the capacity for cognitive functioning. The primary obstacle to cognitive capacity may be attributed to the very unpredictable character of human behavior and mental processes. This posed a hurdle in establishing a genuinely collaborative atmosphere. The cultural specificity and continual evolution of human cognitive activity provide a significant hurdle for AI technology vendors in the robotics sector that want to produce a flawless AI system. The incorporation of artificial intelligence (AI) into collaborative robots may provide additional difficulties for programmers due to the increasing intricacies involved in job execution. The intelligence level of robotic systems is governed by a range of machine learning methods. The absence of a consensus on the acceptability of strong AI comprehension is considered to be attributed to the prevalence of machine learning and deep learning techniques. Collaborative robots provide the capability to automate jobs that are characterized by large volume and repetition, so enabling people to allocate their efforts towards more intellectually demanding projects.

According to Fukuda and Kubota [11], an intelligent robotic perception system refers to a system that provides a collaborative robot with the capability to observe, interpret, and engage in reasoning processes pertaining to its surrounding habitat. The primary factors of an intelligent system of robotic perception, as outlined [12], consist of ML algorithms, environment modeling, and sensors. The components are shown in **Fig 3**. In alignment with the assertions made by Cebollada, Payá, Flores, and Payá [13], it is acknowledged that the current state of artificial intelligence in robotics does not exhibit a significant degree of advancement, particularly in terms of strong AI. Consequently, the following portion of this study will concentrate on the domain of machine learning. The perception system of intelligent robotics plays a crucial role in enabling collaborative robots to make informed judgments, devise plans, and function effectively in real-world settings. This system encompasses a wide range of functions and activities, including but not limited to occupancy grid mapping and object identification. The majority of robotic vision systems use ML methodologies, including both traditional and deep learning methodologies. Machine learning techniques used in robotic perception include several approaches, including

DLNN, supervised classifiers using handcrafted features, unsupervised learning, and perhaps a fusion of numerous methodologies.

According to the findings presented by Semeraro, Griffiths, and Cangelosi [14], the primary focus of ML research is on the information that is received from the robot sensors. The provided dataset contains information that need consolidation in order to facilitate processing using ML techniques. The implementation of ML necessitates data processing, which is contingent upon the specific job at hand. The depiction of the world via sensors is a crucial component of an intelligent system of robotic perception. The current exposition encompasses the model of metric, enabling the representation of the habitat. ML is used at several stages throughout this process.

III. METHODOLOGY

This research used a schematic and systematic approach to examine the progress made in robotic perception systems, with a specific emphasis on deep learning, transfer learning, and fusion methods. The study included the gathering, analysis, and implementation of several late fusion techniques to improve object categorization in a robotic perception environment.

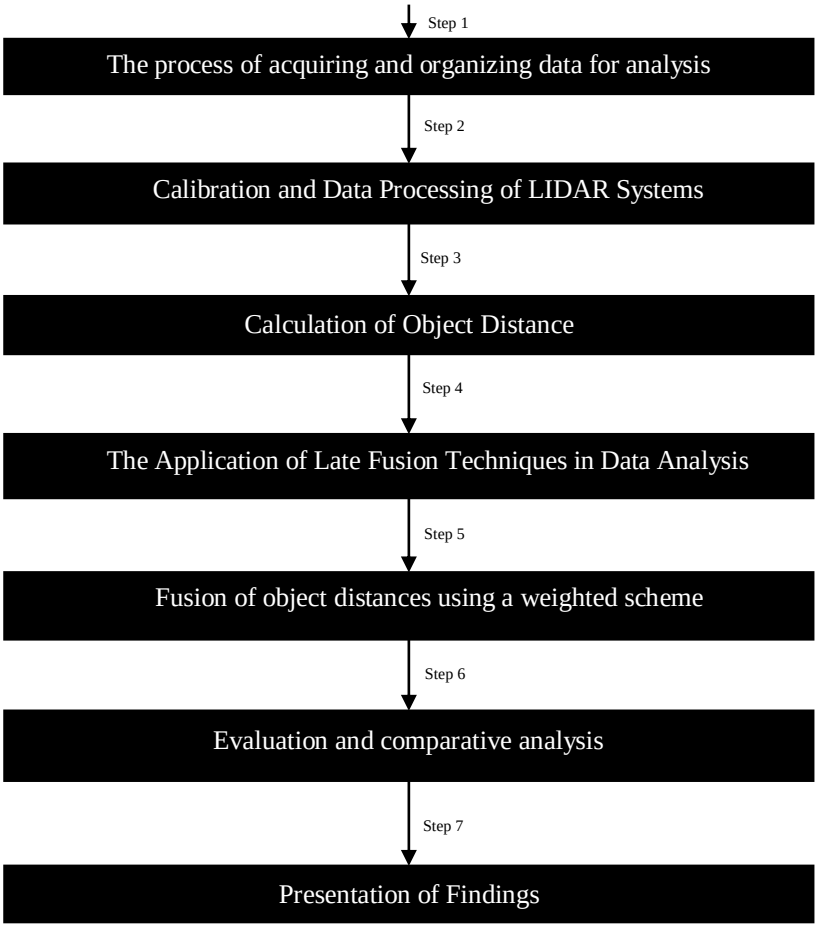


Fig 4. A Systematic Representation of The Methodology.

This study was fortified by conducting a comprehensive analysis of reliable research databases. The search was done using databases such as ScienceDirect, IEEE Xplore and SpringerLink, which encompassed a comprehensive gathering of scholarly articles, conference papers, and peer-reviewed publications that cover advancements in robotic perception systems with special interest on deep learning, fusion techniques and transfer learning. The search criteria included significant terms such as “transfer learning,” “deep learning,” “fusion strategies,” and “robotic perception.” In this case, the present study is founded upon a meticulously selected compilation of at least 25 scholarly sources published recently, prominent publications as well as peer-reviewed papers. Hence, the papers have been one of the sources of information for the development and understanding of different methodologies, challenges and advances in relation to robotic perception technologies today.

The acquisition of RGB images, DM and 3D PCs was the first step using a robotic perception system. To construct a comprehensive classification dataset, entities such as automobiles, people, and bicycles were meticulously removed from RGB images, depth maps, and point clouds by manual means. The dataset was then partitioned into validation, testing, and training subsets, establishing the basis for further analysis. The calibration of the LIDAR system with respect to a camera was of utmost importance in determining the accurate correspondence between 3D coordinates and pixel values on the image plane. The generation of high-resolution two-dimensional representations of three-dimensional point clouds was achieved

by projecting them onto a two-dimensional plane. In this study, a modified version of the Bilateral Filter algorithm was used to generate depth maps, specifically considering the range LIDAR data.

The distances of the objects were estimated using projections of LIDAR, and subsequently, the distance for each point on the DM was computed. The process of augmenting the 3D training dataset included the augmentation of points inside the object point sets, hence enhancing the training of the PointNet model. Several late fusion tactics were used, such as element-wise summation, weighted averaging, bilinear product, and deterministic late fusion processes (average, normalized, minimum, maximum product). Late fusion incorporates machine learning technologies like Support Vector Machines and Genetic technologies. The researchers in this work have introduced a novel late fusion methodology known as Weighted Average with Range (WAR) [15], which incorporates the distances of depth maps and point clouds.

The utilization of the Weighted Average with Range (WAR) approach is a significant improvement since it incorporates a distance-based weighting mechanism to boost the effectiveness of CNNs. The weighting function incorporated the distance of the object as acquired from the LIDAR sensor. The calculation of F-scores considered the quantity of objects at varying distances within both the training and validation datasets. The purpose of this study was to ascertain the weights allocated to different modalities, which were contingent upon the distance measure and the classifier's performance. The classification performance of the different modalities (RGB, DM, PC) on the testing set was evaluated using the F-score metric. The present study aimed to assess the enhancement in classification performance through a comparative analysis of outcomes generated by several deterministic late fusion procedures.

The results were presented in both tabular and graphical formats, demonstrating the effectiveness of late fusion methods and the proposed Weighted Average with Range (WAR) technique. A thorough investigation was undertaken to assess the impact of object distance on late fusion approaches and the overall classification performance. The result of the study implies prospective avenues for future research, emphasizing the need of performing more investigations on the impact of object distance considerations in the fusion of different classifiers. The exploration of potential enhancements or modifications to the methodology has revealed prospects for broader utilization in the field of robotic perception systems. The aforementioned finding has laid the foundation for subsequent examination and exploration inside this significant progressing field. **Fig 4** presents a systematic and schematic representation of the methodology followed to compose the results of this research.

IV. ADVANCEMENTS IN ROBOTIC PERCEPTION SYSTEMS: DEEP LEARNING, TRANSFER LEARNING, AND FUSION TECHNIQUES

Robotic perception systems often include methods that are capable of extracting valuable features or information from the examined data. These methods often include deep learning techniques that use convolutional principles for image processing, as well as approaches for handling 3D data. The notion of image categorization is well defined via the use of networks that include convolutions layers. However, many network systems are expansive and need substantial amounts of time and memory resources. One potential solution to address the challenges of reducing time and memory in machine learning is the implementation of transfer learning or 'neural implants'. These neural implants consist of additional layers that are joined to a pre-trained system, enabling the network to acquire new capabilities with few training samples. The use of 3D point clouds in neural networks is possible without the need of projecting them onto a 2D plane. The PointNet methodology is utilized to perform segmentation, classification, and detection tasks on stationary point clouds, as demonstrated in this study.

On the other hand, both PointFlowNet [16] and FlowNet3D [17] networks demonstrate the capacity to effectively forecast scene flow with precision. These networks possess the ability to accurately predict the three-dimensional displacements of point clouds inside a dynamic environment. In network applications involving point clouds, it is imperative to conduct a thorough analysis of the reliability of the object recognition process in identifying the individual points that comprise the object. This suggests that system or network should possess the capacity to precisely recognize PCs generated by adversaries, hence assuring the network's resilience against hostile attacks. To ensure the promotion of road safety for all persons, encompassing both drivers and non-drivers, it is imperative that the technologies incorporated inside autonomous vehicles encompass precise determination of the vehicle's position and orientation. The research undertaken by [18] presented an architectural framework that integrates several point clouds and deep neural networks (DNN) to address the localization challenges in autonomous driving. The achievement of this outcome is facilitated by the use of eigenvalue computations employing Point Net and 3DCNN. The technique initially identifies the salient features by employing the eigenvalues of adjacent three-dimensional points. The Point Net algorithm is used to extract features, which then serve as the inputs for 3D convolutional neural networks (3DCNN) [19].

The 3D convolutional neural network (3DCNN) applies regularization techniques to the volume across its dimensions. Moreover, recurrent neural networks are used for the purpose of analyzing temporal motion dynamics. One potential approach for processing LIDAR data involves converting the three-dimensional data into a two-dimensional representation. This conversion has the potential to enhance and streamline the use of advanced deep convolutional neural network (DCNN) models. The use of depth and reflectance data allows for the generation of 2D-LIDAR “images” that can be readily analyzed using commercially available convolutional neural networks (CNNs). However, it should be noted that the point clouds produced by the LIDAR sensor exhibit sparsity. Consequently, in order to acquire range maps with a high level of detail, it becomes necessary to sample these points. Various sizes of sliding windows and up sampling techniques, including horizontal disparity processing, Bilateral Filter, Delaunay triangulation, Ordinary Kriging, and Inverse Distance Weighting, may be used to get these maps. Late and early systems of fusion include the integration of output and input data, with the

aim of achieving an improved and more resilient outcome. According to Furuta, Wild, Weng, and Weiss [20], early fusion refers to the integration of data at the level of input of classifiers. An example of early fusion is the combination of images acquired from several modalities, such as depth maps and RGB images, which may be inputted into a single convolutional neural network model with many channels.

In contrast, late fusion involves the aggregation of scores or confidence levels from many learning models at the decision level. Fusion systems may also be implemented using ML algorithms, including SVM, ANN, and other related approaches. In practice, the classification models may be executed concurrently, and at a particular point, the ensuing independent outputs can be merged to facilitate the process.

V. RESULTS AND DISCUSSION

Once the LIDAR system is correctly calibrated using the camera, it becomes possible to build a direct correspondence between the 3D and the associated values of pixel in the image plane. In particular, every LIDAR point will have the positional data represented by pixel coordinates (u, v), in addition to the corresponding range or distance value (r_i), where i varies from 1 to n. To get a two-dimensional representation of the three-dimensional point cloud PC, the projections of PC onto the two-dimensional plane are increased in resolution. In our work, a modified Bilateral Filter version, which is a spatial filtering approach, was employed. The aforementioned outcome was attained by the utilization of a sliding window technique, employing a mask size of $13x13$. In this work, the DM is constructed using range data of LIDAR, while the camera images are solely utilized for visualization and calibration purposes. The Bilateral Filter that has been built employs a tailored weighting methodology to approximate the necessary central pixel 'depth' value within the mask.

Objects and Dataset Distance Distribution

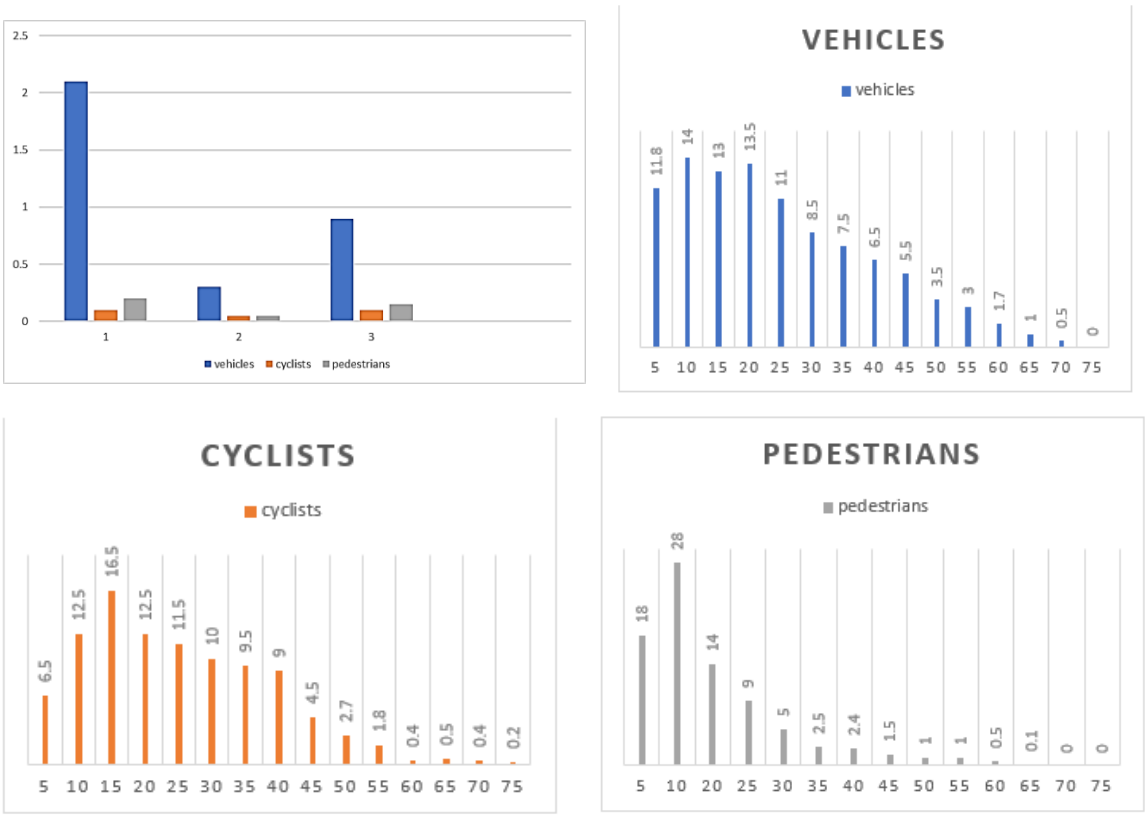


Fig 5. Distribution of Instances Per Class.

The objects in RGB images, DM, and PC were manually extracted, resulting in the creation of a classification dataset including three distinct categories: vehicles (including trucks, automobiles, and vans), pedestrians, and bicycles. **Fig 5** displays the quantities of objects, specifically vehicles, pedestrians, and cyclists, present in the training, validation, and testing sets. The training set contains 20,632 vehicles, 2,827 pedestrians, and 1,025 cyclists. Similarly, the validation set consists of 2,293 vehicles, 314 pedestrians, and 114 cyclists. Lastly, the testing set encompasses 9,825 vehicles, 1,346 pedestrians, and 488 cyclists. Additionally, **Fig 5** also illustrates the objects distribution in relation to their respective distances. One of the primary aims of this study is to assess late fusion strategies that include the object interval as a significant characteristic.

Consequently, the interval-distribution plays a crucial role in this evaluation. The information presented in **Fig 5** depicts the distribution of objects, classified by their respective classes, in correlation with the interval measured in meters as captured by LIDAR. The distances of each item were established through the utilization of projections of LIDAR and by computing the unbiased each point mean interval on the DM. The exclusion of the top and lowest values linked with each object facilitated the achievement of this outcome. Before performing distance calculations on PCs, the 3D dataset of training was analyzed by hypering the quantity of points inside the point set of the objects. The augmentation was implemented to boost the training process of the model of Point Net, which necessitates a consistent input size. Next, we conducted down sampling and up sampling techniques to ensure a consistent input dimension, meaning that each point set associated with an object has an equal amount of points.

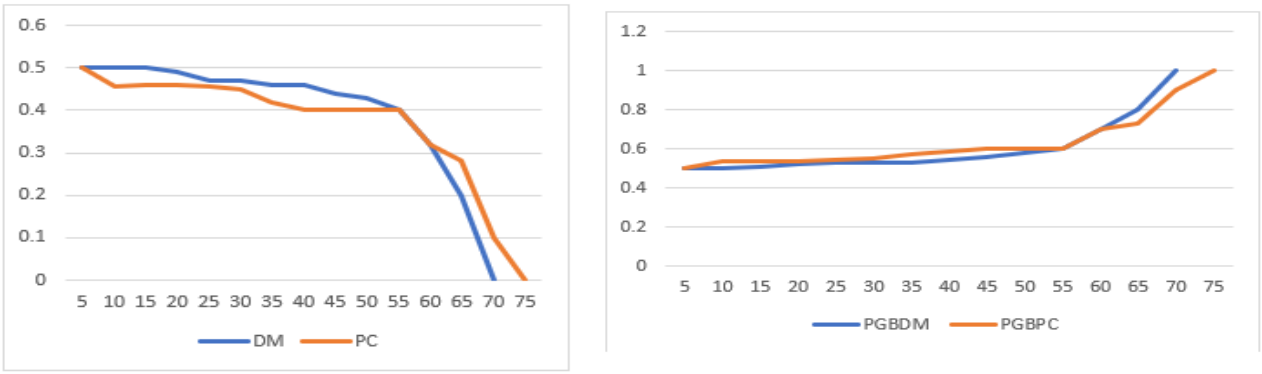
Fig 5, (graph in the top-left corner), illustrates the distribution of instances per class (pedestrians, cyclists, and vehicles) throughout the testing, validation, and training datasets. In **Fig 5**, the training dataset consists of 24,484 objects, the validation dataset contains 2,721 objects, and the testing dataset includes 11,659 objects. The following graphs illustrate the distribution of cases based on their categorization and the measured interval in meters.

Table 1. A Classification Outcomes on The Training Dataset

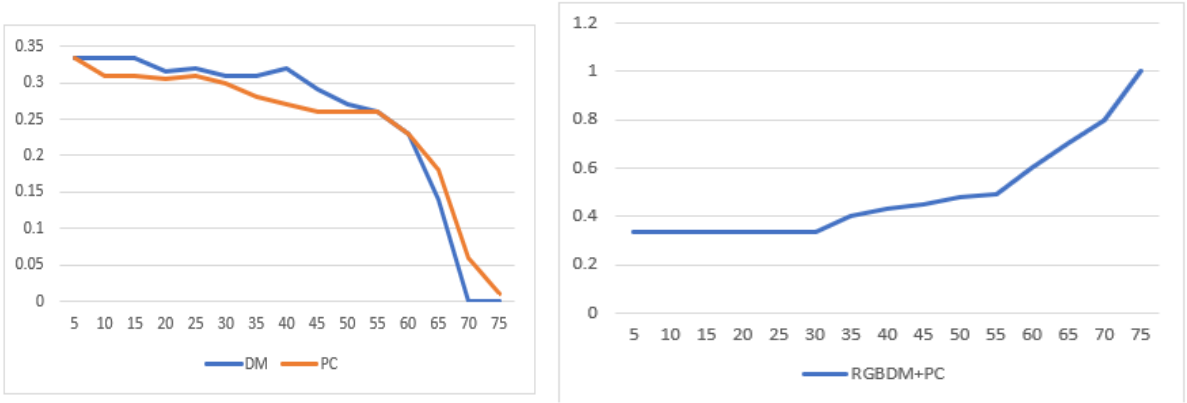
Classes	**PC**				
	64	128	256	512	1024
Pedestrian	75.65	86.85	**95.70**	91.93	86.10
Vehicles	96.65	98.42	**99.41**	99.01	98.04
Cyclists	16.90	72.63	**91.74**	82.41	70.43
Average	63.00	85.97	**96.62**	91.12	84.86

In **Table 1,** the classification results on the training dataset are expressed as F-scores in percentage, for the 3D PCs using the Point Net model.

Weighted Object Distance Fusion



(a) Y_{PC}, Y_C, and Y_{DM}, Y_C modalities



(b) Y_{PC}, Y_{DM}, and Y_C modalities

Fig 6. F-Scores Derived from The Model Of LIDAR, Specifically Using DM And PC, As A Function of Increasing Object Distance.

This research introduces a novel approach to late fusion, which involves using a weighted average (w) based on the distances of the PCs and DMs on the validation and training. The impetus for this study arises from the observation that the performance of the deep-models of LIDAR exhibits a decline as the object`s distance hypers. In contrast, the classification performance of the RGB model which exhibits a very consistent pattern over varying distances from objects. The F-score, a commonly used metric, was computed by taking into account the number of objects at varying distances on both the validation and training sets. This was determined using LIDAR measurements, as seen in **Fig 6**. Consequently, the weighting approach may be seen as a function that depends on the distances of the objects and the classifiers.

The curves shown on the left side of the graph illustrate the weights assigned to the DMs (Deep Models) and PCs (Point Clouds) modalities. Conversely, the right side curves indicate the weight assigned to the RGB (Red Green Blue), specifically denoted as PC-Point Net and wi.DM-CNN models. These weights are directly correlated with the models performance, measured by the F-score multiplied by the Distance metrics. **Fig 6** (a) depicts the normalized average F-score, which attains a maximum value of 0.5. Consequently, it can be inferred that the RGB model outputs (y_C) would possess a maximum weight of 0.5.

On the other hand, **Fig 6** (b) has been standardized at 0.333 (highest value). The weights are contingent using PCs and DMs, as well as their performance on the validation and training sets, which is assessed by the F-score over varying distances of objects. The output (y) of the latefusion technique, referred to as AWR, is constructed based on the following formulation.

$$y = (1 - \sum_i w_i) y_C + \sum_i w_i y L_i \tag{1}$$

The resultant score (denoted as y) following the process of fusion is described as follows: y_C signifies the score of classification obtained from the camera model, yL_i represents the output derived from LIDAR, and i represents the index that represents the classifier of LIDAR, which can be either DM, PC, or a mix of both. The weight, denoted as w_i, associated with a certain classifier of LIDAR, exhibits a relationship with the F-score curve as seen in **Fig 6**. Additionally, this weight is influenced by the distance to the object.

Late Fusion Techniques

The process of late fusion [21] involves combining the predictions from each individual unimodal stream in order to get a final forecast. Fusion may be achieved by several methods, including element-wise summing, weighted average, bilinear product, or a more advanced rank minimization technique. An alternative method for late fusion involves the use of attention mechanisms to choose the most suitable expert for each input signal. The technique proposed by Arévalo, Solorio, Montes-Y-Gómez, and González [22], known as gated multimodal units (GMU), expands upon the existing approach by including gating mechanisms at intermediate feature levels.

In a recent study, Hu, Wang, Nie, and Li [23] have introduced a dense multimodal intermediate (DMI) fusion networking system that facilitates hierarchical joint feature training. The dense fusion operators described in [24] make the assumption that the spatial dimensions of distinct streams are similar, which is also seen in [25]. The use of these approaches in our research is limited to the layers where the spatial dimensions of multimodal variables align, or to following stages of the networking system where spatial dimension has already been integrated. The squeeze operation, as elucidated in this study, facilitates the integration of modalities possessing distinct spatial dimensions at various levels within the hierarchy of the feature.

The late fusion approaches often make the assumption of independence regarding the outputs of the classifiers. In this study, we provide the comparative outcomes obtained via the use of deterministic late fusion procedures, namely normalized, maximum, average, and minimum product.

$$S_{prod} = \frac{\prod_{i=1}^{n} S_i}{\prod_{i=1}^{n}(1 - S_i) + \prod_{i=1}^{n} S_i} \tag{2}$$

In this context, 'n' represents the quantity of models, whereas 'S_i' is the confidence score, sometimes referred to as the output, obtained from a specific model, such as a CNN. Learning methodologies using a SVM and a GA, have also been included. Furthermore, the subsequent techniques were integrated into the measurement of the object's range or distance. This was achieved by using the representations provided by the PCs and/or DMs, which served as an extra feature in conjunction with the scores received from the individual models of Point Net and CNNs. In this particular scenario, the approaches are denoted as GAR and SVMR. The GA fitness function is illustrated by Equation 3, with the objective of maximizing the average F-score.

$$y = I_2 \sum_i w_i y L_i + I_1 \left(1 - \sum_i w_i\right) y_C \tag{3}$$

Where I_1 and I_2 represent individuals, sometimes referred to as "chromosomes". The remaining parameters exhibit identical characteristics to those presented in (1). If the evolutionary algorithm does not include distance in its computations, Equation (3) lacks the inclusion of weighting terms w_i.

Table 2. F-Scores of The Pc, Dm, and Rgb Modalities

Modality	**PC**	**DM**	**RGB**
F-Score	88	89	96

Three kinds of datasets, namely RGB, DM, and PC, have been taken into consideration for the purpose of assessment. The classification results for individual modalities (PC, DM, and RGB) on the testing set, evaluated using the F-score metric, are shown in **Table 2**. The Inception V3 Convolutional Neural Network (CNN) was used for RGB images and depth maps (DMs), while Point Net was used for point clouds (PCs). It is fundamental to note that the findings do not include any fusion approach. **Table 2** presents the outcomes achieved by the utilization of late fusion methods, illustrating that the overall classification efficacy surpassed that of the separate modalities. The usual methodologies utilized for late fusion, including minimum, maximum, average, product, SVM, and GA, have exhibited adequate levels of performance. These approaches fail to consider the magnitudes of the distances between objects. The modalities of fusion Y_{PC}, Y_C, and Y_C, Y_{DM}, Y_{PC}, when integrated with GA and N-Product correspondingly, have demonstrated the most superior overall performance for these two modalities.

VI. CONCLUSION AND FUTURE SCOPE

This research provides a comprehensive examination of the use of multiple classifiers combination, using a late fusion method, for the purpose of object classification inside a robotic perception setting. The study utilized several classification methods, such as deep convolutional neural networks (CNNs), to analyze three distinct forms of sensor data representation. These representations included RGB images obtained from a solitary camera, depth maps (often referred to as range views), and 3D PCs acquired from a 3D Light Detection and Ranging sensor. To assess the methodologies, a 3-class classification object has been developed, including the following categories: vehicles (including automobiles, vans, and lorries), pedestrians, and cyclists. One of the primary purposes of this study was to demonstrate the significance of including the distance of the object as an extra cue inside a perception system. This study is centered on the exploration of late fusion procedures for the purpose of combining or fusing the output, namely the likelihoods or confident levels, obtained from neural networks.

A novel approach, referred to as the WAR technique, has been introduced to boost the effectiveness of CNNs on the training set by using a distance-based weighting function. This weighting function considers the object distance as determined by the LIDAR sensor. The suggested WAR approach demonstrated the highest performance for the Y_C, Y_{DM} modality. On the other hand, the normalized product and the genetic algorithm yielded the good results for the Y_C, Y_{PC} and Y_C, Y_{DM}, Y_{PC} modalities, respectively. The current work shows promise and warrants more investigation, especially in relation to the concept of including a performance measure for object distances in the combination of multi-classifiers. Based on the findings pertaining to the fusion and object distances approach outlined in this study, it can be concluded that LIDAR and camera sensors exhibit a complimentary relationship. The fusion of these two modalities has been shown to enhance the overall performance, thereby establishing its relevance in the context of multisensory perception systems.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests.

References

[1]. A. Elfes, "Using occupancy grids for mobile robot perception and navigation," IEEE Computer, vol. 22, no. 6, pp. 46–57, Jun. 1989, doi: 10.1109/2.30720.

[2]. R. Siegwart, I. Nourbakhsh, and D. Scaramuzza, "Introduction to autonomous mobile robots," Choice Reviews Online, vol. 49, no. 03, pp. 49–1492, Nov. 2011, doi: 10.5860/choice.49-1492.

[3]. I. Kim et al., "Nanophotonics for light detection and ranging technology," Nature Nanotechnology, vol. 16, no. 5, pp. 508–524, May 2021, doi: 10.1038/s41565-021-00895-3.

[4]. B. W. Parkinson and J. J. Spilker, Global positioning system : theory and applications. 1996, p. 114. [Online]. Available: https://arc.aiaa.org/doi/pdf/10.2514/5.9781600866388.0000.0000
[5]. M. Sun, Z. Zhao, and X. Ma, "Sensing and Handling Engagement Dynamics in Human-Robot Interaction Involving Peripheral Computing Devices," 2010 5th ACM/IEEE International Conference on Human-Robot Interaction (HRI), May 2017, doi: 10.1145/3025453.3025469.
[6]. N. Zeng, H. Zhang, Y. Chen, B. Chen, and Y. Liu, "Path planning for intelligent robot based on switching local evolutionary PSO algorithm," Assembly Automation, vol. 36, no. 2, pp. 120–126, Apr. 2016, doi: 10.1108/aa-10-2015-079.
[7]. E. Macias-Garcia, D. Galeana-Pérez, J. Medrano-Hermosillo, and E. Bayro–Corrochano, "Multi-stage deep learning perception system for mobile robots," Integrated Computer-aided Engineering, vol. 28, no. 2, pp. 191–205, Mar. 2021, doi: 10.3233/ica-200640.
[8]. C. Dong, "Remote sensing, hydrological modeling and in situ observations in snow cover research: A review," Journal of Hydrology, vol. 561, pp. 573–583, Jun. 2018, doi: 10.1016/j.jhydrol.2018.04.027.
[9]. A. Buosciolo, G. Pesce, and A. Sasso, "New calibration method for position detector for simultaneous measurements of force constants and local viscosity in optical tweezers," Optics Communications, vol. 230, no. 4–6, pp. 357–368, Feb. 2004, doi: 10.1016/j.optcom.2003.11.062.
[10]. I. Nesnas, L. Fesq, and R. Volpe, "Autonomy for space robots: past, present, and future," Current Robotics Reports, vol. 2, no. 3, pp. 251–263, Jun. 2021, doi: 10.1007/s43154-021-00057-2.
[11]. T. Fukuda and N. Kubota, "An intelligent robotic system based on a fuzzy approach," Proceedings of the IEEE, vol. 87, no. 9, pp. 1448–1470, Jan. 1999, doi: 10.1109/5.784220.
[12]. W. Ma, X. Zhang, and G. Yin, "Design on intelligent perception system for lower limb rehabilitation exoskeleton robot," 2016 13th International Conference on Ubiquitous Robots and Ambient Intelligence (URAI), Aug. 2016, doi: 10.1109/urai.2016.7625785.
[13]. S. Cebollada, L. Payá, M. Flores, and L. Payá, "A state-of-the-art review on mobile robotics tasks using artificial intelligence and visual data," Expert Systems With Applications, vol. 167, p. 114195, Apr. 2021, doi: 10.1016/j.eswa.2020.114195.
[14]. F. Semeraro, A. Griffiths, and A. Cangelosi, "Human–robot collaboration and machine learning: A systematic review of recent research," Robotics and Computer-Integrated Manufacturing, vol. 79, p. 102432, Feb. 2023, doi: 10.1016/j.rcim.2022.102432.
[15]. J. R. Mosig, "The Weighted Averages algorithm revisited," IEEE Transactions on Antennas and Propagation, vol. 60, no. 4, pp. 2011–2018, Apr. 2012, doi: 10.1109/tap.2012.2186244.
[16]. A. Behl, D. Paschalidou, S. Donné, and A. Geiger, "PointFlowNet: Learning Representations for Rigid Motion Estimation From Point Clouds," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, doi: 10.1109/CVPR.2019.00815.
[17]. X. Liu, C. R. Qi, and L. J. Guibas, "FlowNet3D: Learning Scene Flow in 3D Point Clouds," 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2019, doi: 10.1109/CVPR.2019.00062.
[18]. Md. Z. Hussain, M. Ashraf, D. K. Singh, A. Haldorai, D. K. Mishra, and T. N. Shanavas, "Intelligent data post and read data system like to feed for IoT sensors," International Journal of System Assurance Engineering and Management, Jun. 2022, doi: 10.1007/s13198-022-01683-5.
[19]. S. Ji, W. Xu, M. Yang, and K. Yu, "3D convolutional neural networks for human action recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 35, no. 1, pp. 221–231, Jan. 2013, doi: 10.1109/tpami.2012.59.
[20]. R. Furuta, C. Wild, Y. Weng, and C. D. Weiss, "Capture of an early fusion-active conformation of HIV-1 gp41," Nature Structural & Molecular Biology, vol. 5, no. 4, pp. 276–279, Apr. 1998, doi: 10.1038/nsb0498-276.
[21]. C. K. Mohan, N. Dhananjaya, and B. Yegnanarayana, "Video Shot Segmentation Using Late Fusion Technique," 2008 Seventh International Conference on Machine Learning and Applications, Jan. 2008, doi: 10.1109/icmla.2008.88.
[22]. J. Arévalo, T. Solorio, M. Montes-Y-Gómez, and F. A. González, "Gated multimodal units for information fusion," arXiv (Cornell University), Feb. 2017, [Online]. Available: https://arxiv.org/pdf/1702.01992.pdf
[23]. D. Hu, C. Wang, F. Nie, and X. Li, "Dense Multimodal Fusion for Hierarchically Joint Representation," ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), May 2019, doi: 10.1109/icassp.2019.8683898.
[24]. Q. Dai, X. Cheng, Y. Qiao, and Y. Zhang, "Agricultural Pest Super-Resolution and identification with attention enhanced residual and dense fusion generative and adversarial network," IEEE Access, vol. 8, pp. 81943–81959, Jan. 2020, doi: 10.1109/access.2020.2991552.
[25]. V. De Silva, J. Roche, and A. M. Kondoz, "Robust fusion of LiDAR and Wide-Angle camera data for autonomous mobile robots," Sensors, vol. 18, no. 8, p. 2730, Aug. 2018, doi: 10.3390/s18082730.