Computers, Materials & Continua
DOI: 10.32604/cmc.2023.035442
Article

Tech Science Press

A Novel Capability of Object Identification and Recognition Based on Integrated mWMM

M. Zeeshan Sarwar[1], Mohammed Hamad Alatiyyah[2], Ahmad Jalal[1], Mohammad Shorfuzzaman[3], Nawal Alsufyani[3] and Jeongmin Park[4,*]

[1]Department of Computer Science, Air University, Islamabad, Pakistan
[2]Department of Computer Science, College of Sciences and Humanities in Aflaj, Prince Sattam Bin Abdulaziz University, Al-Kharj, Saudi Arabia
[3]Department of Computer Science, College of Computers and Information Technology, Taif University, Taif, 21944, Saudi Arabia
[4]Department of Computer Engineering, Tech University of Korea, 237 Sangidaehak-ro, Siheung-si, Gyeonggi-do, 15073, Korea
*Corresponding Author: Jeongmin Park. Email: jmpark@tukorea.ac.kr

Received: 21 August 2022; Accepted: 04 November 2022

Abstract: In the last decade, there has been remarkable progress in the areas of object detection and recognition due to high-quality color images along with their depth maps provided by RGB-D cameras. They enable artificially intelligent machines to easily detect and recognize objects and make real-time decisions according to the given scenarios. Depth cues can improve the quality of object detection and recognition. The main purpose of this research study to find an optimized way of object detection and identification we propose techniques of object detection using two RGB-D datasets. The proposed methodology extracts image normally from depth maps and then performs clustering using the Modified Watson Mixture Model (mWMM). mWMM is challenging to handle when the quality of the image is not good. Hence, the proposed RGB-D-based system uses depth cues for segmentation with the help of mWMM. Then it extracts multiple features from the segmented images. The selected features are fed to the Artificial Neural Network (ANN) and Convolutional Neural Network (CNN) for detecting objects. We achieved 92.13% of mean accuracy over NYUv1 dataset and 90.00% of mean accuracy for the Redweb_v1 dataset. Finally, their results are compared and the proposed model with CNN outperforms other state-of-the-art methods. The proposed architecture can be used in autonomous cars, traffic monitoring, and sports scenes.

Keywords: Artificial intelligence; convolutional neural network; depth images; interactive-object detection; machine learning

1 Introduction

The majority of current RGB-D datasets were gathered with depth sensors, such as Kinect or LiDAR. Kinect can only be utilised for inside situations, although LiDAR is frequently employed for outside scenes. Due to the variety of situations, it is challenging to achieve decent results in the wild while training on outdoor scene datasets. Accessibility of publicly available network datasets like as ImageNet [1], NYU v1 [2], and Streetview, as well as video classification benchmarking datasets such as Caltech 101 [3] has opened the path for significant advancements in object detection in recent years. Using an RGB-D (Kinect) camera, we have observed the beginning of a new generation of detection systems that are capable of producing exceptional colour and depth pictures. These strategies significantly improve the robots' object detecting capabilities. As extracted features can increase the quality of object identification determination, this article conducts experiments with depth photographs. With the easy availability of RGB-D sensors and datasets, which can able to appropriate colour and depth, it is hoped that occlusion and lighting issues may be successfully resolved. In some circumstances, thin, dark, and foggy conditions might make object identification more difficult. In [4], the authors suggested a methodology that enhanced the effectiveness of 3D feature-based object identification. Utilizing 3D characteristics allows for high precision [5]. In furthermore, the depth of information in such situations is crucial for a various applications, including security monitoring, the medical profession, military, 3D interactive games, autonomous driving, and mapping. A model presented by [6] performs picture categorization using kernel characteristics. The authors of [7] described an architecture that performs localization and object recognition on RGB-D pictures by subtracting the background and segmenting the subjects. These approaches enhance the quality of object identification and recognition; hence, we will conduct studies with these depth pictures. The suggested architecture integrates Convolutional Neural Network (CNN) and convolution filters with form context-specific properties.

To develop robots capable of perceiving the environment as humans do, researchers have focused heavily on Scene Semantic Recognition (SSR), automated analysis of item placements [8] and structural connection between many objects in landscape photos. Improving the dependability of SSR in practical applications such as security navigation to automatically detect suspicious/violent situations [9], recognition of social interaction types in public settings, differentiating between diverse sports scenarios, and remote sensing still faces significant obstacles.

One difficulty encountered in 3D image processing is the mathematically consistent blending of various point clouds gathered from multiple angles. Various object detection techniques have been proposed for RGB-D photos and movies. Some systems employ the technique of object localization based on depth maps, such as establishing a Conditional Random Field (CRF) model and a system to comprehend interior scenes. Utilizing 3D characteristics yields outcomes with high precision. Proposed is a method that extracts the fusion of characteristics such as depth edges, 3D forms, and size features. Using this fusion, they were able to obtain considerable high performance with RGB-D pictures [10]. Image normal estimate is conceptually comparable to fitting a plane to a local point cloud in the three-dimensional space [11]. Using a multilayer perceptron for scene comprehension, [12] proposes a system based on a hybrid Histogram of Gradients (HOG) and local geometrical characteristics for multi-object identification and recognition. Another research [13] detects circles using the Hough transform. Using fundamental visual descriptors such as form, colour, and roughness in conjunction with their feature fusion for object recognition during pattern recognition. In a separate work, the author [14] described an efficient shape-matching procedure that makes use of form contexts. The similarity between the shapes of the target objects is estimated by locating a transformation between shape points.

A approach based on multi-object categorization is proposed to conduct scene classification on a variety of benchmark datasets in order to circumvent the issues inherent in scene classification. The suggested approach initially preprocesses the photos. In the second stage, the improved Watson Mixture Model (mWMM) technique is used to generate efficient segmentation results, and clustering is conducted. Multiple characteristics are retrieved in the third stage, including 3D-point clouds, form features, and a bag of words. In the final phase, the characteristics are provided to two distinct architectures, Artificial Neural Network (ANN) and CNN [15] in order to identify photos from two difficult datasets. The following are the key contributions of this research:

- An approach for multi-object detection and scene understanding based on modified WMM, ANN, and CNN (Vgg-16) is proposed.
- Improved segmentation for the detection of multiple regions of different objects using modified WMM and 3D-geometric features are the main contributions of this work.
- Novel 3D-geometric features for scene understanding have refined the scene recognition accuracy with both ANN and CNN architectures.
- The proposed model's efficiency and effectiveness are validated with two different publically available datasets.
- Other state-of-the-art approaches are compared to the suggested method's outcomes. Section 2 discusses relevant work. Section 3 describes the approach and suggested scene categorization system in depth. The fourth section provides an analysis of the experimental outcomes and a comprehensive explanation of the information. Section 5 contains the paper's summary.

2 Related Work

Our technique is connected to a large body of work on both CNNs for fusion and machine vision. In furthermore, we briefly analyses the appropriateness of CNN's detailed estimate. It is beyond the goal of this research to do a comprehensive literature review of CNNs for these three parameters; therefore, we will present a brief summary of the existing studies, with a focus on more recent publications.

2.1 CNN for Fusion

Reference [16] proposes a system based on convolutional and nonlinear machine learning for learning and classifying RGB-D image features. Iteratively artificial neural network retrieve high-level characteristic features, whereas convolutional layers extract low-level characteristics. This is updated by [17], which recommends a tractor trailer methodology that employs fewer labelled data but achieves comparable results to the state-of-the-art. Another publication [18] illustrates how to employ CNNs for fusing diverse gyroscope inputs, including infrared and RGB images. The results of these studies suggest that the effectiveness of object person identification can be improved by combining the data from both camcorders, as appeared differently in relation to simply employing one camera.

Early fusion [19], advanced fusion [16,18], and late blending [18,20] are the three basic classifications of existing fusion techniques according to the level of data indirection engaged for combining. In early fusion, also renowned as image blending, the raw sensor readings are combined to create fused data prior to the application of information extraction techniques. CNN performance in feature extraction, depth estimation, and eye shadow appreciation has been significantly improved by pixel-level fusion approaches [21]. Intermediate fusion, also referred to as feature fusion, consolidates each raw data's extracted features. In late combining, also recognised as decision-level fusion, the scanners are considered individually to each sensor, and their results are then combined to determine the final detection. For instance, [22] devised a late combination for amalgamating depth details and RGB relevant information to improve the performance of object characterization. This approach to information fusion includes two CNN infrastructures for each data modality and incorporates the highlights only after the connections have been run. Our proposed fusion architectural style actually creates a depth information from RGB as opposed to an RGB-D sensor like the Microsoft Kinect.

2.2 CNN Detection

It has been demonstrated that CNN-based techniques perform better than exquisite methodology such as HOG [23] and SIFT [24]. In [25], the use of Region-based Convolutional Neural Networks (R-CNN) significantly enhanced the accuracy of pattern recognition. It identifies area specific recommendations (i.e., regions of significance that are likely to have particles) before classifying these provinces as different classifiers or backgrounds using a CNN. R-CNN has the limitation of individually calculating the CNN for each receptive field, which is both time-and electricity. To improve performance and scalability, faster R-CNN [26] omits the judicious search strategy for actually creating instrument region guidelines.

The significant proportion of CNN-based detection equipment consist of one-stage processes (e.g., SSD [27], YOLO [28]) and method adopted (e.g., R-CNN [25], Fast/Faster R-CNN [26,29], R-FCN [30]). Two-stage detection methods are slightly slower than separate detection equipment because a the outside module is needed to create targetable positions. Nevertheless, their classification performance is enhanced as a result of the stringent example consideration. The challenging instances are those for which the model makes poor predictions. In comparision, yet another object detection methods create a congested specimens of possible classifier is based more quickly and directly by omitting the second per-region categorization and merely predicting anchor boxes and associated model is trained. None of eachother evaluated consolidating multiple sources of data to improve detection and recognition.

2.3 CNN for Depth Estimation

Attributed to the reason that mostght pictures include numerous bounding boxes, extensive textural variations, and intricate geometric elements, adjusted is a significant obstacle in image interpretation. Diverse depth estimation strategies employing supervised [31,32] and unsupervised [33–35] learner methods have been developed to address this problem. Recent supervised learning techniques [31,32] use CNNs for 3d reconstruction to avoid manually-crafted object functionalities and computationally demanding test-time optimization. Also discussed in DeMon [31] is thickness learning utilising stereo data. Using two unconstrained frames, their method produces reliable depth predictions. This approach utilises a variety of supervisory techniques, including depth and optical flow.

CNNs exhibit promise effectiveness for this task, but supervised techniques require costly and time-consuming datasets with extensive labelling. Various studies apply a personality learning approach to estimate depth maps from unlabeled video sequences in order to overcome this issue. Self-supervised learning approaches overcome the difficulty of background subtraction by educating a network to predict the appearance of a target image from the perspective from another image. Df-net [33] is an independent pedagogical approach for simultaneously training depth estimation and optical flow estimation [34]. It generates 2D gaussian filter by backprojecting the produced 3D scenario flow utilising predicted scene depth and camera motion to generate 2D optical flow. SfMLearner [35] is a later part instructional strategy that use monocular video sequence only for training. It discovers depth by utilising the geometric relationship amongst depth and cameraman position.

3 Proposed Methodology

In this segment, the suggested architecture for object detection is discussed. Fig. 1 demonstrates a general summary of the proposed architecture. Image normal is acquired from depth image and then clustering is performed using the Watson mixture model to assist the segmentation phase. After that, scale-dependent 3D geometric features are computed along with the color features that will be used to detect objects in the later phase of the architecture. Finally; for object detection purposes, ANN and CNN are applied.

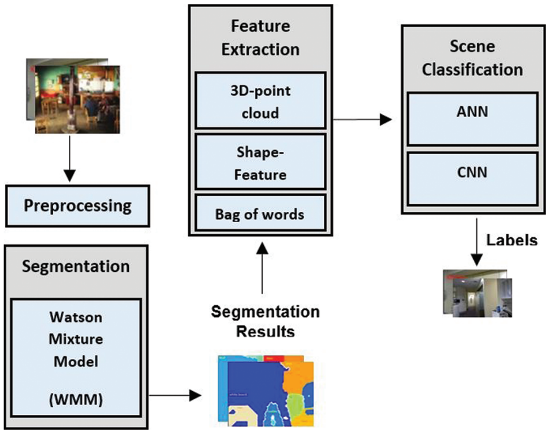


Figure 1: Block diagram of the proposed system for multiple objects recognition

3.1 Data Acquisition and Pre-Processing

A number of researchers have proposed surface normal in their recent work. In-depth images, image normal are the unit vectors having 3D properties that draw the positioning of the pixels. The most collective method used to compute normal was the plane fitting method [7]. In [36] a method was proposed by repeating shape patterns and appearance primitives in indoor RGB-D image data, and compared those primitives to new images to acquire a normal map against depth images. Normally image normal acquiring is considered to be a computationally expensive process because of 3D point cloud fitting. Since normal space selection is helpful for point cloud and image normal registration, different approaches for estimating surface normal from a 3D point cloud have been projected in the nonfiction. In the proposed approach, depth maps have been used to perform clustering using the image normal technique. 3D points (J, K, L) in the camera synchronize arrangement are projected onto a pixel (j, k) as from carrying a depth image to normal clustering space

$$x = f_L^J + C_j \tag{1}$$

$$y = f_L^K + C_k \tag{2}$$

where f is focal length, (C_j, C_k) is the optical midpoint of the depth camera, and they are acquired during the camera calibration procedure. Then, a 3D point (J, K, L) is parameterized as a function of a pixel (j, k).

$$J\left(x,y\right)=\frac{x-C_{j}}{f}L\left(x,y\right) \tag{3}$$

$$K\left(x,y\right)=\frac{x-C\_k}{f}L\left(x,y\right) \tag{4}$$

With a depth sensor, $L\ (x,\ y)$ can be obtained as a pixel assessment on a depth image. Reminder that the measure of L and f is calibrated in millimeters beforehand.

The sequence diagram presented in Fig. 2 provides a visual illustration of this concept. Furthermore, gradients of the profundity picture are obtained, and a normal tensor is generated for each pixel based on the differences. Perfectly natural averaging is then used to a satellite image derived from the depth picture.

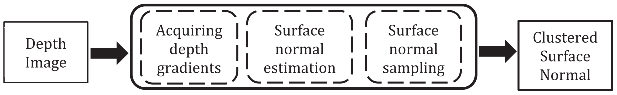


Figure 2: Image normal extraction through depth image

3.2 *Data Segmentation*

Throughout data classification, the subsurface normal of distance pictures is computed. Normals are produced for each pixels by choosing neighboring pixels within a depth minimum and estimating a least-squares surface. Then, updated WMM is used to cluster the normal. Each constellation in the output is a collection of pixels from the same location. Figs. 3a and 3b illustrate picture normals *vs.* depth images (d).

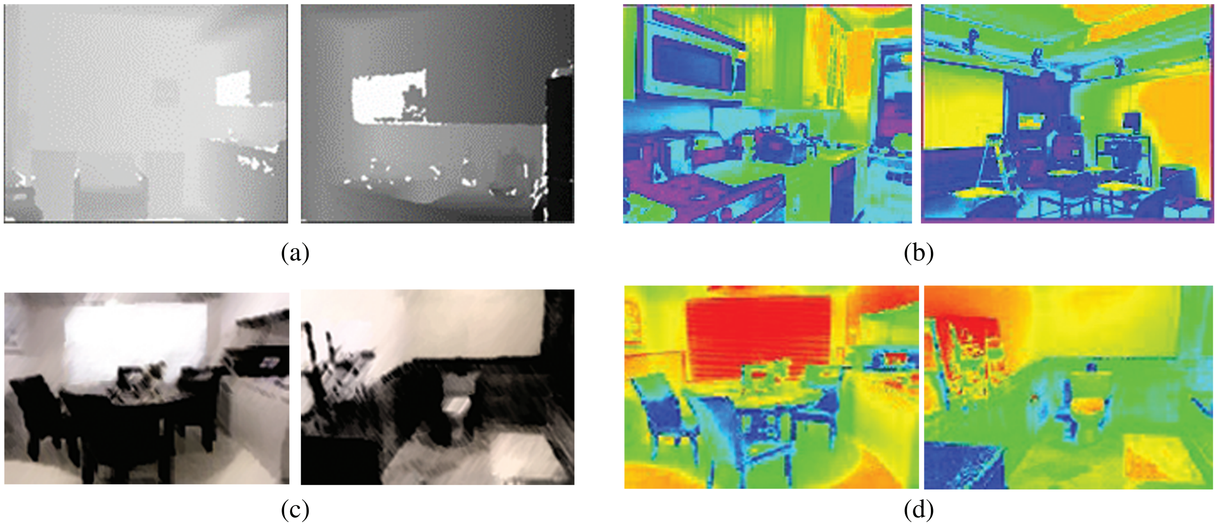
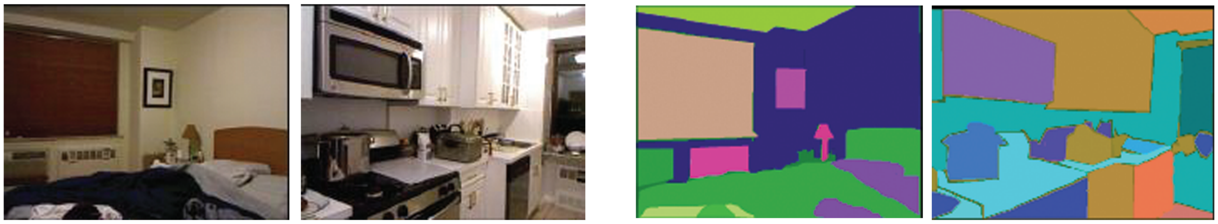


Figure 3: Segmentation approach. (a) Depth map from NYU V1; (b) acquired image normal against NYU V1 depth map; (c) depth map from RedWeb V1; (d) acquired image normal RedWeb V1 depth map

3.3 Modified Watson Mixture Model (WMM)

Modified Watson Mixture Model is a productive model, which undertakes that the data models are issued from a combination of multivariate Watson distributions (mWDs) [37]. Usually, Watson distribution is used for modeling data. However, in the proposed organization, it is used to achieve segmentation and to assist the multi-object detection phase. In [18] researchers proposed an architecture that begins with the Bregman Soft Clustering technique for the mWMM.

After soft clustering, a set of mWMMs using hierarchical collective clustering is produced. In the end, it applies a model selection method to select the optimal mWMM is applied. In the proposed model, more than one mWMM is generated against a single depth map and an optimum mWMM is picked gives the best clusters. The proposed technique holds distributional information acquired from depth gradients and needs no repetitive numerical calculation during the optimization procedure. Moreover, the approach provides a lower certain on the peripheral likelihood. Consequences are shown in Figs. 4a and 4b. The first two images, show RGB representation of those depth images against which Watson distribution was applied. The next two images are the results of the segmentation process.



(a) (b)

Figure 4: Multiple objects segmentation using mWMM based on depth information (a) represents the RGB object image (b) represents selected mWMM segmentation

The segmentation step starts with the initialization of the parameters that are variational in nature. The optimization of the variational subsequent distribution involves a series of optimization. Firstly, we use the current distributions over the model parameters to examine the responsibilities. Next, these obligations are engaged to re-estimate the subsequent distribution over the parameters. The process is guaranteed to converge as the lower bound. A summary of the process is presented in Algorithm 1.

Algorithm 1: mWMM feature extraction for multi-object segmentation

```
Input:    Depth image with extracted image normals
Output:    Segmented multi object image with different color representation
          Concatenated human-object feature vectors as $\{V_1, V_2, V_3 \ldots\ldots\ldots V_O\}$
          //initiating feature descriptors//
        Concatenated set the number of ingredients, G. ← [ n]
      Set the prior distribution parameters: ← [ n]
for i = 1:M do
Initialize $n_k$ using the K-means algorithm.
              //extracting image normal features //
Optimization of the subsequent Distribution
Update repository $n_k$ by computing $a_{0,\ k}$ $b_{0,\ k}$
Update subsequent distribution parameters $\alpha_k$, $\beta_k$
```

(Continued)

Algorithm 1: Continued

```
    Update $\lambda^{-}_{k}$ by setting $\lambda^{-}_{k} = a_k\ b_k$
    **end**
    Evaluate lower bound.
                //The lower bound is monitored in each iteration until convergence//
  **end**
```

3.4 Features Descriptors

There are various feature extraction techniques including spatiotemporal motion variation features [38], hybrid features approach [39], ECG and GMM features [40], body joint features [41], and Ridge body parts features [42]. We have used three different feature descriptors. Those include geometric features, shape features, and a bag of words features.

3.4.1 Geometric Features

Some geometric features that show a given 3D lattice model lie on the model's surface. 3D geometric information assumes a significant part in numerous issues identified with computer vision applications.

By that, as it may, their scale-dependent nature, such as the relative variation in the spatial degrees of nearby geometric constructions. For this reason, a scale-space sort of portrayal is fabricated that dependably encodes the scale fluctuation of its surface geometry. The given geometry is addressed with its surface normal and a thick and ordinary 2D plane of it is processed by defining the surface on a 2D plane. At this point, a scale-space of this surface typical field is worked by determining and applying a scale-space administrator that effectively represents the geodesic distances on a superficial level. A 2D representation of the 3D geometry is given as a 3D lattice model by first opening up the outside of the model onto a 2D plane. A significant arrangement of scale-subordinate features can be procured from the subsequent typical space portrayal.

Geometric edges and sharp focuses are extracted at various scales. To set up these edges, the first- and second-request subsidiaries of the depth map are gotten attentively. The outcome is a bunch of scale-subordinate 3D geometric features that give a rich and extraordinary reason for the exhibition of 3D unique basis. In Fig. 6, there are some visual instances of geometrical focuses procured over the depth properties of the given pictures.

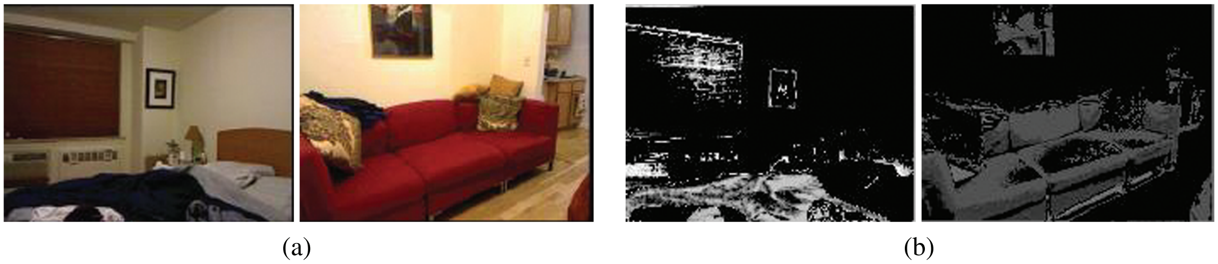
$$\varepsilon\left(i\right) = \frac{1}{\left|y\left(i\right)\right|}\sum\nolimits_{\mathrm{v}\,\in\,\mathrm{y(i)}}\frac{\left|\left|i\ -\ v\right|\right|}{\left|\left|\varnothing\left(i\right) - \varnothing\left(v\right)\right|\right|} \tag{5}$$

3.4.2 Shape Features

In feature extraction, specific geometric forms, such as cylindrical, rectangular, and other configurations, are used to extract significant features from various features in an image. In the suggested approach, form total number of positive on the profundity qualities of the material shapes are utilized. From depth information, it can be tried by fitting a quantized model based on boundaries, shape priors, and detected spatial properties. A contour-based approach often has minute applicability to specific shapes, such as generalized cylinder shapes. In some scenarios objects in a messy environment are turned into region of interest (ROI) parts. In this section, a straight-line strategy is applied to detect the shape of our interest regions. First of all, region contour is extracted by the boundary acquiring technique.

$$\|w_D\| = \sqrt{(}\left(x1_i - x2_j\right)2 + \left(y1_i - y2\right)2\big) \tag{6}$$

where $\boldsymbol{w_D}$ represents the distance and *(x1i, y1j*) and (*x2i, y2j*) are the *x*, *y* synchronizes of first and additional points, correspondingly. We can see some visuals demonstration computed over gradient properties in Figs. 5a and 5b.



(a)

(b)

Figure 5: Shape features extraction over NYU V1 Dataset (a) RGB image and (b) shape features extracted

3.4.3 *Bag of Word Features*

A bag of words representation of the image features has also been used for better scene categorization. Since this algorithm does not account for spatial relationships between the features, it is bound to miss categorize some scenes.

- Starting with the extraction of features from the training set, These features lead us to develop a vocabulary that will help in image classification.
- The next step is to cluster all the features found in images and then to find a difference between features vocabulary by using the center of the cluster.
- After taking the feature extracted from images, categorize each feature as the word it is closest to it in the vocabulary. In this way, a bag of words representation is made.
- Against each bag of words, a histogram is constructed.

$$SD = \sum\nolimits_{m=1}^{k} \sum\nolimits_{x\in zi} (x - zi)^2 \tag{7}$$

$$A^*, B^* = arg\frac{1}{M}\sum\nolimits_{F=i}^{N}\sum\nolimits_{g=i}^{k} B_{F,G}\left(A_i - x_j\right)^2 \tag{8}$$

3.5 *Objects Classification*

Two different methods have been used for multi-object recognition and classification. ANN and CNN. Both are robust in multi object scenario and scene classification problems.

3.5.1 *Artificial Neural Network (ANN)*

ANN is a computer model used for modelling statistical data across non-linear data. It is a tool for computer science that is informed by the nervous system and replicates the actual brain's learning system to execute learning. ANN discovers various data correlations or input-output correlations utilizing artificially generated neurons. Fig. 6 depicts an ANN with input, production, and one maybe more convolutional units. Hidden layers are used to turn the input sequence into the activation function. Numerous methods may be utilized to create ANN; in the proposed method, a nutrient form multi-layer perceptron (MLP) approach is employed to recognize different object characteristics.

Algorithm 2: ANN processing flow

```
Input:   Extracted feature vector from image
Output:  Labeled multi object classification
         Concatenated human-object feature vectors as {$k_1$, $k_2$ ........ $k_O$ }
         // Initializing Stop criteria//
         Set the prior distribution parameters:
   for while = stop criteria
   Calculates ep (w) for each pattern
   e1:$\sum_{p=1}^{P}$ $e_p$ (w)T $e_p$ (w);
   Calculates Jp (w) for each pattern
   repeat
   Calculates $\Delta$w
   e2: = $\sum_{p=1}^{P}$ $e_p$ (w + $\Delta$w)T $e_p$ (w + $\Delta$w)
   if (e1 <= e2) then
   $\mu := \mu * \beta$.
     Endif
     end
     Until (e2 < e1);
     $\mu := \mu/\beta$
     w: = w + $\Delta$w
     endwhile
```

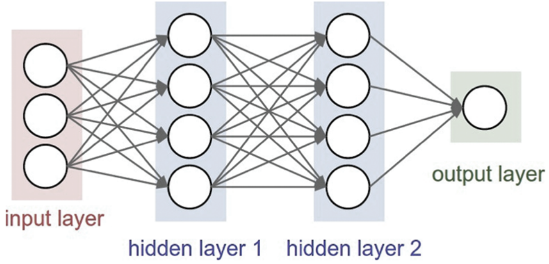


Figure 6: Block illustration of the artificial neural network for multiple matters appreciation

3.5.2 Convolutional Network (CNN)

A pre-trained CNN VGG-16 classical is also used for object classification in the proposed system. Fig. 7 shows that the input to the first convolutional layer of VGG-16. VGG-16 comprises 16 convolutional covers and three fully associated layers that depict the depth of the complex which leads the model toward high accuracy classification. The proposed approach uses VGG-model only for classification purposes with the help of our pre-acquired features like a bag of features, 3D geometric and shape features.

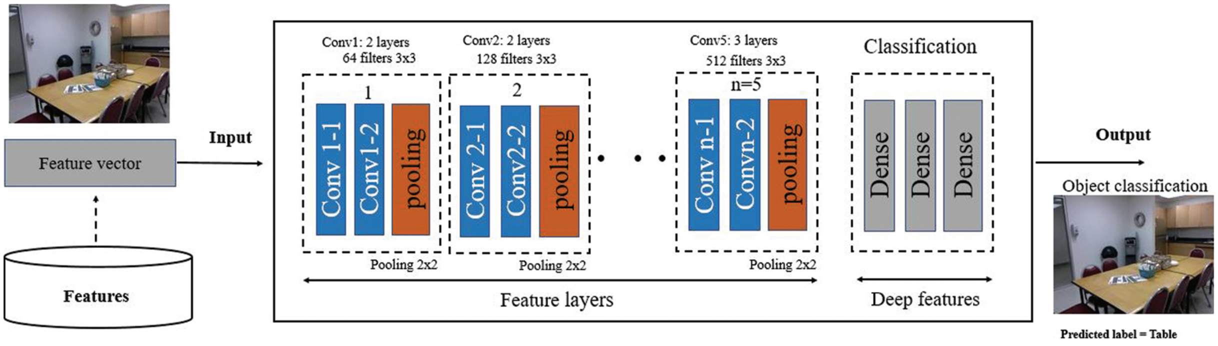


Figure 7: Sequence of steps multiple objects recognition over an image from NYU V1 dataset using the convolutional neural network

4 Experimental Setup and Results

This section describes the study's preparation and assessment procedure in details.

4.1 Dataset Descriptions

Two different datasets have been used to test the proposed methodology. These datasets comprise various scenes with multiple objects and various classes. A description of these datasets is given in the next section.

4.1.1 ReDWeb Dataset

ReDWeb V1 is a comprehensive database comprised of a variety of photos and their respective complex comparable depth maps. The ReDWeb V1 [2] dataset includes 3,600 RGB-D photos of residential and commercial settings. This collection includes synchronised RGB-D frames between Kinect v2 and Zed photogrammetry. We build differential maps and used an appropriate stereoscopic matching approach before converting them with calibration settings. Also included is a per-pixel accuracy map of discrepancy. Footage are taken in several locations, such as workplaces, rooms, dormitories, exposition centres, streets, and roads. The collection includes 200 distinct scenarios with varied items. We chose 500 crisp photos. Here, 450 photos are utilised for train, while the other 50 photographs are synthesised for testing. We evaluated the reliability of our transparency evaluation system predicated on NYUv1 and DIML. Fig. 8 depicts genuine RedWeb v1 photos.

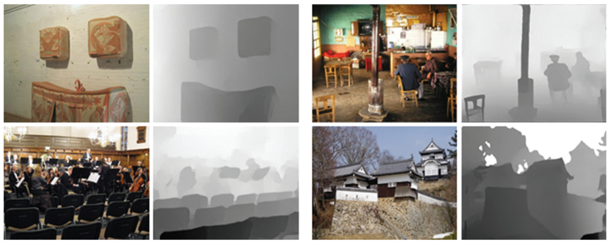


Figure 8: A set of sample images from the RedWebV1 dataset with corresponding depth images

4.1.2 NYU V1 Dataset

The NYU V1 [4] data set comprises of image sequences collected by the Deep and RGB cameras of the Motion Capture system from a range of interior situations. It consists of seven distinct scenario types with a minimum of sixty-four scenes. The NYU v1 collection contains many perspectives on the same and distinct items, as well as RGB and complexity photos [4]. It comprises 64 distinct indoor sceneries, a variety of daily things viewed from various perspectives, and over a thousand distinct classes. To determine the effectiveness of the product recognition and scene classification challenge, the geometry attribute, 3D-points collection, and characteristics were utilised to train Random Forest (RF) and neural network-based classifiers. RF was educated on the chosen features and ANN was trained on a combination of both features. The experiments took place on a stochastic training and testing combination of 65% and 35% for laboratory investigation. Selected RGB and depth pictures from the NYU v1 Datasets are displayed in Fig. 9.

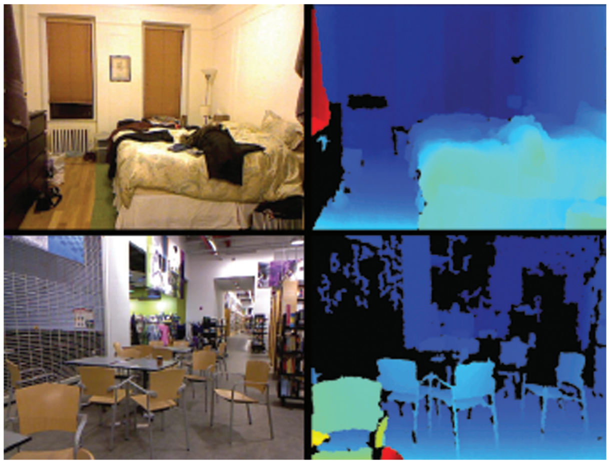


Figure 9: A set of sample images from the NYU V1 dataset with corresponding depth images

4.2 Experimental Results

In this part, the setup and assessment are explained in greater detail. In the evolutionary sense, precision of classification and comparability with established state approaches were tested by analysing all indoor photos. Because of the strong object segmentations (mWMM), which exhibits greater efficiency in object identification utilising ANN and Network architectures, the suggested system produced consistent results.

4.2.1 Experiment 1: Using RedWeb V1 Dataset

Considering the RedWeb V1 dataset, the proposed system was applied for scene classification accuracy. Tables 1 and 2 show that the major object classes of the considered dataset produce significant performance in terms of accuracy.

Table 1: Confusion matrix of individual object class accuracies over the Redweb v1 dataset using ANN

Objects	WH	CF	CR	CHR	CMP	MD	LIB	LAB	BS	COR
WH	**0.87**	0.03	0.01	0.01	0.01	0.02	0.03	0.01	0.00	0.01
CF	0.01	**0.89**	0.02	0.01	0.01	0.02	0.00	0.01	0.02	0.01
CR	0.02	0.01	**0.88**	0.01	0.02	0.01	0.01	0.02	0.01	0.01
CHR	0.02	0.01	0.02	**0.87**	0.01	0.00	0.01	0.02	0.03	0.01
CMP	0.01	0.01	0.02	0.00	**0.89**	0.01	0.02	0.01	0.02	0.01
MD	0.02	0.01	0.01	0.01	0.02	**0.86**	0.02	0.01	0.03	0.01
LIB	0.03	0.02	0.01	0.02	0.01	0.01	**0.84**	0.03	0.01	0.02
LAB	0.02	0.01	0.02	0.00	0.01	0.02	0.01	**0.87**	0.02	0.02
BS	0.01	0.02	0.01	0.01	0.03	0.02	0.03	0.01	**0.85**	0.01
COR	0.03	0.01	0.01	0.02	0.01	0.01	0.01	0.02	0.01	**0.87**

Note: WH = ware house, CR = computer room, CH = chair, CMP = computer, MD = mobile device, LAB = laboratory, COR = corridor.

Table 2: Confusion matrix of detailed object class accuracies over the Redweb v1 dataset using CNN

Objects	WH	CF	CR	CHR	CMP	MD	LIB	LAB	BS	COR
WH	**0.91**	0.00	0.02	0.01	0.02	0.01	0.01	0.00	0.01	0.01
CF	0.01	**0.90**	0.02	0.01	0.01	0.00	0.02	0.01	0.01	0.01
CR	0.02	0.01	**0.89**	0.01	0.02	0.01	0.01	0.01	0.00	0.02
CHR	0.01	0.00	0.01	**0.92**	0.02	0.01	0.02	0.00	0.00	0.01
CMP	0.02	0.01	0.01	0.01	**0.88**	0.01	0.01	0.02	0.03	0.01
MD	0.01	0.02	0.01	0.00	0.00	**0.92**	0.01	0.02	0.01	0.01
LIB	0.01	0.02	0.00	0.01	0.01	0.01	**0.90**	0.02	0.01	0.01
LAB	0.03	0.01	0.02	0.01	0.01	0.01	0.02	**0.87**	0.01	0.01
BS	0.01	0.02	0.01	0.01	0.01	0.01	0.02	0.01	**0.90**	0.00
COR	0.02	0.01	0.01	0.02	0.01	0.01	0.02	0.01	0.01	**0.88**

Note: WH = ware house, CR = computer room, CH = chair, CMP = computer, MD = mobile device, LAB = laboratory, COR = corridor.

4.2.2 Experimentation: Using the NYU V1 Dataset

Throughout experiments with the NYU V1 dataset, classification accuracy score of 89.1% as shown in Table 3 using CNN and 78% using ANN shown in Table 4.

Table 3: Confusion matrix of specific entity class precisions over the NYUv1 dataset using ANN

Objects	BR	BD	BS	CF	KIT	LR	OFF
BR	**0.87**	0.02	0.02	0.01	0.03	0.02	0.03

(Continued)

Table 3: Continued

Objects	BR	BD	BS	CF	KIT	LR	OFF
BD	0.03	**0.85**	0.02	0.02	0.03	0.02	0.03
BS	0.02	0.02	**0.88**	0.03	0.02	0.01	0.02
CF	0.03	0.04	0.03	**0.84**	0.01	0.02	0.03
KIT	0.03	0.01	0.01	0.02	**0.87**	0.03	0.03
LR	0.03	0.02	0.02	0.03	0.03	**0.85**	0.02
OFF	0.04	0.02	0.03	0.01	0.02	0.02	**0.86**

Note: BR = bar room, BD = bed room, BS = bicycle, KIT =kitchen, LR = laboratory, OFF = office.

Table 4: Confusion matrix of individual object class accuracies over the NYUv1 dataset using CNN

Objects	BR	BD	BS	CF	KIT	LR	OFF
BR	**0.91**	0.01	0.02	0.03	0.02	0.01	0.02
BD	0.01	**0.92**	0.02	0.01	0.01	0.01	0.02
BS	0.02	0.01	**0.92**	0.01	0.02	0.01	0.01
CF	0.01	0.02	0.01	**0.92**	0.01	0.01	0.02
KIT	0.02	0.01	0.02	0.01	**0.90**	0.03	0.01
LR	0.03	0.01	0.01	0.02	0.03	**0.89**	0.01
OFF	0.01	0.02	0.03	0.01	0.00	0.02	**0.91**

Note: BR = bar room, BD = bed room, BS = bicycle, KIT = kitchen, LR = laboratory, OFF = office.

Table 5 illustrations the evaluation of the proposed prototypical with some state-of-the-art methods using equally NYU V1 and Redweb V1 dataset.

Table 5: Assessment of Redweb V1 dataset with state of the art technique

Methods	NYU V1	Redweb V1
	Mean accuracy	Mean accuracy
K. Chen [4]	–	60.5
S. Gupta [5]	–	65.0
A. Zeng et al. [6]	–	78.1
Silberman et al. [2]	70	–
Multiscale convnet [3]	51.1	–
Proposed ANN Approach	**88.0**	**89.1**
Proposed CNN Approach	**92.13**	**90.0**

The classification results of the three classifiers, i.e., Random forest, Artificial Neural Network (ANN), and Convolutional Neural Network on Nyuv1, and Redweb V1 datasets are reported in Tables 6 and 7. All three classifiers were trained on the training customary. The classification outcomes in Table 6 were obtained by using the testing set. As shown in Table 7, the classification effects of the suggested structure are better for both datasets as they have higher F-measures, Precision, and Recall scores than those obtained with other classifiers. The overall results showed that the proposed method using CNN achieved better performance than other state-of-the-art methods.

Table 6: The classification results of three classifiers on the Redweb V1 dataset

Multi Objects	Random Forest			ANN			CNN		
	Precision	Recall	F-Measures	Precision	Recall	F-Measures	Precision	Recall	F-Measures
WH	0.781	0.770	0.762	0.879	0.862	0.874	0.891	0.890	0.885
CF	0.750	0.751	0.759	0.864	0.859	0.854	0.889	0.890	0.885
CR	0.756	0.749	0.753	0.880	0.872	0.880	0.883	0.880	0.890
CHR	0.752	0.751	0.750	0.879	0.866	0.876	0.878	0.875	0.872
CMP	0.739	0.745	0.749	0.880	0.878	0.878	0.890	0.885	0.880
MR	0.748	0.740	0.742	0.879	0.874	0.875	0.878	0.875	0.870
LIB	0.749	0.748	0.7750	0.880	0.879	0.879	0.889	0.880	0.887

Note: WH = ware house, CR = computer room, CH = chair, CMP = computer, MD = mobile device, LAB = laboratory, COR = corridor.

Table 7: The classification results of three classifiers on the NYU V1 dataset

Multi Objects	Random Forest			ANN			CNN		
	Precision	Recall	F-Measures	Precision	Recall	F-Measures	Precision	Recall	F-Measures
BR	0.840	0.839	0.843	0.890	0.890	0.889	0.911	0.910	0.899
BD	0.850	0.848	0.847	0.889	0.887	0.886	0.899	0.900	0.886
BS	0.843	0.840	0.842	0.890	0.886	0.888	0.901	0.892	0.898
CF	0.851	0.850	0.849	0.889	0.887	0.890	0.910	0.910	0.900
KIT	0.849	0.850	0.846	0.890	0.888	0.882	0.900	0.887	0.889
LR	0.851	0.847	0.845	0.890	0.884	0.890	0.910	0.889	0.899

5 Conclusion

In this paper, a novel and effective approach for the segmentation and classification of single and heterogeneous objects is provided. Particles were segmented using the powerful algorithm Watson Hybrid Concept. Furthermore, many characteristics were extracted from both collections. The suggested scheme outperforms previous state-of-the-art technologies in terms of computational, segments outcomes, and precision, as determined by experiments performed. In future consideration, the scholars want to conduct an in-depth analysis of photos of outdoor space in order to increase the accuracy of semantic segmentation and discover a solution to the computational effort of semantic segmentation.

Funding Statement: This study was funded by the MSIT (Ministry of Science and ICT) of Korea through the ITRC (Information Technology Research Center) funding programme (IITP-2023-2018-0-01426) under the direction of the IITP (Institute for Information & Communications Technology Planning and Evaluation).

Conflicts of Interest: The authors declare that they have no conflicts of interest to report regarding the present study.

References

[1] B. C. Russell, A. Torralba, K. P. Murphy and W. T. Freeman, “LabelMe: A database and web-based tool for image annotation,” *International Journal of Computer Vision*, vol. 77, no. 3, pp. 157–173, 2008.

[2] N. Silberman, D. Hoiem, P. Kohli and R. Fergus, “Indoor segmentation and support inference from rgbd images,” in *Proc. Int. Conf. on European Conf. on Computer Vision,* Florence Italy, pp. 746–760, 2012.

[3] J. Ma, L. Zheng, Y. Yaguchi, M. Dong and R. Oka, “Image classification based on segmentation-free object recognition,” in *Proc. Int. Conf. on Image Processing*, Hong Kong, China, pp. 2157–2160, 2010.

[4] K. Chen, D. Zhang, L. Yao, B. Guo, Z. Yu *et al.,* “Deep learning for sensor-based human activity recognition: Overview, challenges, and opportunities,” *ACM Computing Surveys*, vol. 54, no. 4, pp. 1–40, 2021.

[5] S. Gupta, R. Girshick, P. Arbelez and J. Malik, “Learning rich features from rgb-d images for object detection and segmentation,” in *Proc. Int. Conf. on European Conf. on Computer Vision*, Zurich, Switzerland, pp. 345–360, 2014.

[6] A. Zeng, S. Song, M. Nießner, M. Fisher, J. Xiao *et al.,* “3dmatch: Learning local geometric descriptors from rgb-d reconstructions,” in *Proc. Int. Conf. on IEEE Conf. on Computer Vision and Pattern Recognition*, Hawaii, USA, pp. 1802–1811, 2017.

[7] J. Fan and T. W. S. Chow, “Exactly robust kernel principal component analysis,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 3, pp. 749–761, 2020.

[8] A. Jalal, M. Z. Sarwar and K. Kim, “RGB-D images for objects recognition using 3D point clouds and RANSAC plane fitting,” in *Proc. Int. Conf. on Applied Science and Technology*, Bhurban, Pakistan, pp. 518–523, 2021.

[9] L. Fei-Fei, R. Fergus and P. Perona, “One-shot learning of object categories,” *IEEE Transection on Pattern Analysis on Machine Intelligence*, vol. 28, no. 4, pp. 594–611, 2006.

[10] Y. Cheng, R. Cai, Z. Li, X. Zhao and K. Huang, “Locality-sensitive deconvolution networks with gated fusion for rgb-d indoor semantic segmentation,” in *Proc. Int. Conf. on Computer Vision Pattern Recognition*, Hawaii, USA, pp. 1475–1483, 2017.

[11] C. Farabet, C. Couprie, L. Najman and Y. LeCun, “Learning hierarchical features for scene labeling,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 8, pp. 1915–1929, 2012.

[12] I. Akhter, A. Jalal and K. Kim, “Pose estimation and detection for event recognition using sense-aware features and adaboost classifier,” in *Proc. of. Conf. on Applied Sciences and Technologies (IBCAST)*, Islamabad, Pakistan, pp. 500–505, 2021.

[13] Y. Ghadi, I. Akhter, M. Alarfaj, A. Jalal and K. Kim, “Syntactic model-based human body 3D reconstruction and event classification via association based features mining and deep learning,” *PeerJ Compututer Science*, vol. 7, pp. e764, 2021.

[14] I. Akhter, “Automated posture analysis of gait event detection aia a hierarchical optimization algorithm and pseudo 2D stick-model,” M.S. Thesis, Dept. Computer science, Air University, Islamabad, Pakistan, 2020.

[15] J. Zhang, C. Tsoi and S. Lo, "Scale invariant feature transform flow trajectory approach with applications to human action recognition," in *Proc. Int. Joint Conf. on Neural Networks*, Beijing, China, pp. 1197–1204, 2014.

[16] R. Socher, B. Huval, B. Bath, C. Manning and A. Ng, "Convolutional-recursive deep learning for 3d object classification," in *Proc. Advances in Neural Information Processing Systems*, Harrsha and Harveys, Lake Tahoe, pp. 656–664, 2012.

[17] Y. Cheng, X. Zhao, K. Huang and T. Tan, "Semi-supervised learning and feature evaluation for rgb-d object recognition," *Computer Vision and Image Understanding*, vol. 139, pp. 149–160, 2015.

[18] F. Farahnakian, J. Poikonen, M. Laurinen and J. Heikkonen, "Deep convolutional neural network-based fusion of RGB and IR images in marine environment," in *Proc. Intelligent Transportation Systems Conf.*, Auckland, New Zealand, pp. 21–26, 2019.

[19] Z. Cai, Q. Fan, R. Feris and N. Vasconcelos, "A unified multi-scale deep convolutional neural network for fast object detection," in *Proc. European Conf. on Computer Vision*, Amsterdam, Netherlands, pp. 1–10, 2016.

[20] F. Farahnakian, J. Poikonen, M. Laurinen and J. Heikkonen, "Deep convolutional neural network-based fusion of rgb and ir images in marine environment," in *Proc. IEEE Intelligent Transportation Systems Conf.*, Auckland, New Zealand, pp. 21–26, 2019.

[21] D. Xu, W. Wang, H. Tang, H. Liu, N. Sebe *et al.,* "Structured attention guided convolutional neural fields for monocular depth estimation," in *Proc. Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 1–6, 2018.

[22] M. Schwarz, H. Schulz and S. Behnke, "Rgb-d object recog-nition and pose estimation based on pre-trained convolutional neural network features," in *Proc. Int. Conf. on Robotics and Automation*, Seattle, WA, USA, pp. 1329–1335, 2015.

[23] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *Proc. Computer Vision and Pattern Recgnition*, San Diego, CA, USA, pp. 886–893, 2005.

[24] D. Lowe, "Distinctive image features from scale-invariant key-points," *International Journal Computer Vision*, vol. 60, pp. 91–110, 2004.

[25] R. Girshick, J. Donahue, T. Darrell and J. Malik, "Region-based convolutional networks for accurate object detection and segmentation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, pp. 142–158, 2016.

[26] S. Ren, K. He, R. Girshick and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," *Advances in Neural Information Processing Systems*, vol. 28, pp. 91–99, 2015.

[27] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed *et al.,* "Ssd: Single shot multibox detector," in *Proc. European Conf. on Computer Vision*, Amsterdam, The Netherlands, pp. 21–37, 2016.

[28] J. Redmon, S. Divvala, R. Girshick and A. Farhadi, "You only look once: Unified, real-time object detection," in *Proc. on Computer Vision and Pattern Recognition*, Salt Lake City, UT, USA, pp. 779–788, 2016.

[29] R. Girshick, "Fast r-cnn," in *Proc. on Computer Vision*, Washington, DC, USA, pp. 1440–1448, 2015.

[30] J. Dai, Y. Li, K. He and J. Sun, "R-Fcn: Object detection via region-based fully convolutional networks," *Advances in Neural Information Processing Systems*, vol. 29, pp. 379–387, 2016.

[31] B. Ummenhofer, H. Zhou, J. Uhrig, N. Mayer, E. Ilg *et al.,* "Demon: Depth and motion network for learning monocular stereo," in *Proc. Computer Vision and Pattern Recognition*, Honolulu, HI, USA, pp. 456–464, 2016.

[32] B. Li, Y. Dai and M. He, "Monocular depth estimation with hierarchical fusion of dilated cnns and soft-weighted-sum inference," *Pattern Recognition*, vol. 83, pp. 328–339, 2017.

[33] Y. Zou, Z. Luo and J. Huang, "Df-net: Unsupervised joint learning of depth and flow using cross-task consistency," in *Proc. European Conf. Computer Vision*, Munich, Germany, pp. 1–6, 2018.

[34] C. Godard, O. Aodha and G. Brostow, "Digging into self-supervised monocular depth estimation," in *Proc. Computer Vision*, Seoul, Korea, pp. 1–6, 2019.

[35] T. Zhou, M. Brown, N. Snavely and D. Lowe, "Unsupervised learning of depth and ego-motion from video," in *Proc. Computer Vision and Pattern Recognition*, Honolulu, Hawaii, pp. 234–240, 2017.

[36] L. Di, C. Guangyong, C. Daniel, H. P. -Ann and H. Hui, "Cascaded feature network for semantic segmentation of rgb-d images," in *Proc. Int. Conf. on Computer Vision*, Venice, Italy, pp. 1320–1328, 2017.

[37] A. Jalal, S. Kamal, A. Farooq and D. Kim, "A spatiotemporal motion variation features extraction approach for human tracking and pose-based action recognition," in *Proc. ICIE)*, Fukuoka, Japan, pp. 1–6, 2015.

[38] Y. Y. Ghadi, I. Akhter, S. A. Alsuhibany, T. al Shloul, A. Jalal *et al.,* "Multiple events detection using context-intelligence features," *Intell. Autom.\& Soft Comput*, vol. 34, no. 3, pp. 1455–1471, 2022.

[39] I. Akhter, A. Jalal and K. Kim, "Adaptive pose estimation for gait event detection using context-aware model and hierarchical optimization," *J. Electr. Eng.\& Technol.*, vol. 310, pp. 1–9, 2021.

[40] A. Jalal, M. Batool and S. Tahir, "Markerless sensors for physical health monitoring system using ecg and gmm feature extraction," in *Proc. on Applied Sciences and Technologies*, Islamabad, Pakistan, pp. 340–345, 2021.

[41] A. Jalal, S. Kamal and D. Kim, "Depth map-based human activity tracking and recognition using body joints features and self-organized map," in *Proc. on Computing, Communications and Networking Technologies*, Hefei, China, pp. 1–6, 2014.

[42] A. Jalal, Y. Kim and D. Kim, "Ridge body parts features for human pose estimation and recognition from RGB-D video data," in *Proc. of Fifth Int. Conf. on Computing, Communications and Networking Technologies (ICCCNT)*, Hefei, China, pp. 1–6, 2014.