

A Multimodal Data Processing System for LiDAR-Based Human Activity Recognition

Jamie Roche¹, Varuna De-Silva, Joosep Hook, Mirco Moencks², and Ahmet Kondo, *Senior Member, IEEE*

Abstract—Increasingly, the task of detecting and recognizing the actions of a human has been delegated to some form of neural network processing camera or wearable sensor data. Due to the degree to which the camera can be affected by lighting and wearable sensors scantiness, neither one modality can capture the required data to perform the task confidently. That being the case, range sensors, like light detection and ranging (LiDAR), can complement the process to perceive the environment more robustly. Most recently, researchers have been exploring ways to apply convolutional neural networks to 3-D data. These methods typically rely on a single modality and cannot draw on information from complementing sensor streams to improve accuracy. This article proposes a framework to tackle human activity recognition by leveraging the benefits of sensor fusion and multimodal machine learning. Given both RGB and point cloud data, our method describes the activities being performed by subjects using regions with a convolutional neural network (R-CNN) and a 3-D modified Fisher vector network. Evaluated on a custom captured multimodal dataset demonstrates that the model outputs remarkably accurate human activity classification (90%). Furthermore, this framework can be used for sports analytics, understanding social behavior, surveillance, and perhaps most notably by autonomous vehicles (AVs) to data-driven decision-making policies in urban areas and indoor environments.

Index Terms—Convolutional neural network, faster RCNN, Fisher vector, human activity recognition (HAR), multimodal machine learning (ML).

I. INTRODUCTION

HUMAN activity recognition (HAR) has engaged many to become one of the most popular and active research fields in machine vision. Motivated by the academic value and practical applications in sports analytics [1], human-machine interface [2], ambient-assisted living (AAL) [3], and autonomous vehicle (AV) research [4], [5], HAR is characterized by the actions taken by individuals in a group, or individuals acting alone. If a machine is to understand the actions taken by people, it needs to develop an intelligence that will allow it to reason about its behavior.

Manuscript received 10 August 2020; revised 8 December 2020 and 22 April 2021; accepted 23 May 2021. Date of publication 24 June 2021; date of current version 16 September 2022. This article was recommended by Associate Editor A. Li. (*Corresponding author: Jamie Roche.*)

Jamie Roche, Varuna De-Silva, Joosep Hook, and Ahmet Kondo are with the Institute for Digital Technologies, Loughborough University London, London E15 2GZ, U.K. (e-mail: a.j.roche@lboro.ac.uk).

Mirco Moencks is with the Institute for Manufacturing, University of Cambridge, Cambridge CB2 1TN, U.K.

Color versions of one or more figures in this article are available at <https://doi.org/10.1109/TCYB.2021.3085489>.

Digital Object Identifier 10.1109/TCYB.2021.3085489

To reason about the behavior of people, three core challenges must be addressed. The primary challenge is getting a machine to understand the complex interaction performed by people when they modify their behavior quickly depending on circumstances. These interactions are influenced by numerous factors, such as distance between agents or direction of travel. As such, the location and action of agents are equally important in facilitating a machine's ability to understand what people are doing.

The second challenge is the selection of attributes to be measured and the sensors that measure them. Being a popular and active research field, there are many different approaches. As such, sensor inaccuracy, constraints, or obtrusiveness need to be considered. For example, wearable sensors provide reliable information about the actions of an agent. Unfortunately, wearable sensors are not as prolific as they need to be. While smartphone accelerometers are part of the way there, they cannot capture as much information as a camera can. Similarly, the camera is excellent at capturing a wide field of view (FoV) but is prone to interference from lighting conditions. If each sensor has a shortcoming, the inevitable choice is to overcompensate on modality, thus improving the chances of capturing all the necessary data required to improve accuracy.

The third challenge is the sheer quantity of data required to train a classifier. Most contemporary supervised machine-learning (ML) methods require examples to learn. In principle, given unlimited data, most supervised ML paradigms should be able to classify any activity under any circumstances. However, there is a question of whether we can access the required data to achieve this task in practice. To that end, a framework that can utilize the data more intelligently is a significant challenge.

To address these challenges, we proposed a novel CNN architecture to leverage multimodal ML and sensing redundancy benefits. The proposed framework uses a pretrained CNN (ResNet-50) [8] and a region proposal network (RPN) as an object detector [6]. The purpose of the object detector is to identify a region of interest (ROI)—containing a person performing an activity—in an image. The corresponding ROI is translated and aligned to point cloud data before being segmented and classified. This approach extends research on a 3-D modified Fisher vector representation derived from a Gaussian mixture model (GMM) [7].

The proposed framework addresses the three challenges, showing that sensor data redundancy and multimodal ML enable us to use the data more intelligently, accommodate

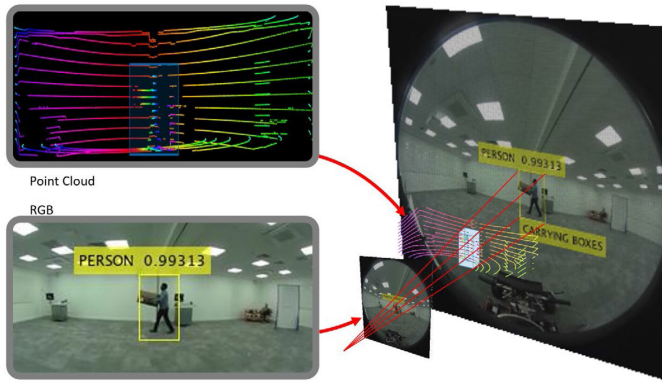


Fig. 1. Proposed HAR pipeline. Given RGB data, we first detect a subject and generate an ROI. Each RGB ROI is translated and aligned onto the 3-D point cloud. The corresponding 3-D ROI is then passed onto the classifier before deciding what activity is being performed.

diverse feature space, and reason jointly about agents' activities and location. Specifically, we developed a technique where a detected ROI is projected onto light detection and ranging (LiDAR) data, followed by a point cloud classification of an activity performed by a human, as depicted in Fig. 1.

The primary contribution of this work is a new, accurate and reliable, method of HAR, which leverages multiple sensing modalities. The secondary and tertiary contributions are the detection of an ROI and the segmentation of subjects captured in 3-D data. The remainder of this article is organized as follows. Section II provides an overview of the current state of the art in HAR. Section III details the proposed framework. Section IV reports on the comparative performance and visual results of the proposed framework. Section V discusses some of the limitations before concluding this article in Section VI.

II. RELATED WORK

The technology and sensors integral to HAR vary to great a degree. This research uses a camera and a LiDAR sensor to classify people performing different activities. The data collected by the camera is regular 2-D data, and the LiDAR collects 3-D or point cloud data. To that end, we use 3-D data and point cloud data interchangeably throughout the course of this article.

Multimodal ML is a research field with one of its earliest applications in the field of audio-visual speech recognition (AVSR) [8], to contemporary fields of emotion recognition [9], [10] and tracking multiple agents [11] or anomalies in traffic [12]. Like sensor data fusion, in the sense that it uses multiple sensor modalities, Multimodal ML uses multiple ML algorithms, either in parallel or in series, to improve a framework [13]. Five core technical challenges need to be addressed in multimodal ML [14]. They can be described as: 1) *representation*, which is concerned with the difference between data types; 2) *translation* which is the mechanism of moving data from one plane to another; 3) *alignment* which indicates the proximal relationship between two different sensor streams; 4) *fusion* which denotes the joining of sensor data; and 5) *co-learning* which means how knowledge is recognized from one modality for use with another.

Despite numerous developments in ML, the majority are demonstrated in tightly controlled settings with well-defined outcomes, typically perform singular tasks and generally use only one sensor stream [15]. Developing agents that make decisions using context information gathered from real-world environments remains an essential task, especially if we want a machine to understand what is going on around them.

The context of a scenario can be captured from various types of instruments, measurement techniques, and sensors. Also known as multimodal sensing, it is where multiple sensor modalities capture context information about the environment they are working in [16] and [17]. Commonly referred to as sensor data fusion, the challenges can be categorized into two groups [5], [16]: 1) challenges at acquisition level and 2) challenges due to the uncertainty of data sources.

One of the most prolific applications of multimodal ML and sensor data fusion is in the area of AAL. Typically used in hospitals, rehabilitation centers, and elderly care units [18], the algorithms are designed to determine what situations require urgent medical assistance. This process is usually done using multiple signals to determine a drift away from homeostasis [19] and identified using multimodal ML by combining multiple biometrics to deduce what has just occurred.

A closely related research field is multimodal ML for HAR and pedestrian detection [20]–[22]. Using the principles derived from AAL, it is expected that researchers could decrease road accidents and improve road safety. For example, Diederichs *et al.* [23] and Rehder *et al.* [24] provided an interesting application of multimodal ML for a pedestrian recognition system that matches the pedestrians predicted intention with the driver's direction and, depending on the actions, brake the vehicle to avoid a collision.

For the most part, research into HAR has focused on either RGB, RGB-D, or wearable sensor data. These techniques extract human silhouettes, tracking them in the temporal domain and classify activities based on patterns analyzed [12] or velocity and acceleration data from an agent to make a classification. For example, research in [25]–[27] used micro-electromechanical systems (MEMS) to facilitate a wearable sensor-based approach to activity recognition. Suitable for both indoor and outdoor environments, their size and lightweight make MEMS ideal for such scenarios. Furthermore, when combining multimodal ML methods in similar manners to researchers in [28], the argument for using wearable sensors becomes abundantly clear. Regrettably, though, MEMS have a long way to go to become ubiquitous. Even though smartphones are commonplace and can be used for this purpose, a single stream of low-level data, such as acceleration data, is insufficient to identify the more nuanced differences between walking with a phone and just walking [20], [29].

In some cases, a single stream of low-level data can be used to detect actions such as falls. However, in reality, the performance of such recognition systems is often significantly lower than the reported research results [30]. For example, Igual *et al.* [31] cross-validated different fall detection algorithms and found unfavorable results. These unfavorable results can be attributed to variations in device

hardware and the relative position of sensors [32]. Clearly, when dealing with accelerometer data, the placement of sensors is paramount [33]–[38]. For example, falling, a desirable activity to recognize, and possible to do so with a single stream of low-level data from a smartphone, the type sensor, and place where the user carries the device is a significant factor influencing the algorithms' ability to determine the action accurately. Although, possible to substitute a smartphone with wearable MEMS. We arrive back at the same problem of scantness, making the optical and range sensors an ideal substitute for the less than prolific wearable MEMS.

Wearable MEMS, such as the electrocardiogram (ECG) and photoplethysmogram (PPG), can provide reasonably reliable information. However, particle applications like HAR suffer attenuation due to clothing artifacts [39], are easily influenced by diet [40] and provide erroneous data due to prescribed medication [41]. For example, artifactual signals caused by clothing can corrupt normal cardiac signals at the electrode. Dietary sodium, associated with unfavorable alterations in left ventricular mass, hypertension, and increased water retention, can significantly lower carotid-femoral pulse wave velocity (PWV). Conversely, medications such as L-Dopa, typically used in the treatment of Parkinsons patients, cause postural hypotension—the pooling of blood in the lower limbs—resulting in an inability to distinguish the diastolic peak of the PPG signal. This is not to say that wearable MEMS are not valuable sensors used to gather data, just that when applied to HAR they are scant and can easily be influenced by external factors, which result in inaccurate or erroneous data.

Herein the reason why some activity recognition researchers are focusing on RGB sensors rather than wearable MEMS. For example, Bagautdinov *et al.* [1] proposed a unified framework for HAR using a recurrent neural network (RNN) to identify human subjects with an imaging sensor. Researchers in [42] used RNNs and imaging data to perform the same task except for single subject activity recognition. And researchers in [43] present a vision-based HAR system using a mobile camera and histogram of oriented gradient (HOG) extracted from the feature space.

Expanding on research using RGB, some scholars have made a further transition over to a data stream that can accommodate depth information and true color. Using two parallel CNNs, called ResNet [44], to classify activities, Mukherjee *et al.* [45] presented a method of capturing the motion information of a whole video by producing a dynamic image corresponding to the input video. Similarly, Cippitelli *et al.* [46], Feng *et al.* [47], and Arzani *et al.* [48] proposed an activity recognition algorithm exploiting skeleton data extracted from RGB-D sensors, an activity scene recognition method based on a 3-D skeleton sequence, and a probabilistic graphical model creating loopy subgraphs applicable for modeling simple and complex activities, respectively.

While most of the activity recognition systems have focused on wearable MEMS, RGB, or RGB-D data, most contemporary systems are equipped with a camera and range sensors such as LiDAR. These sensors often complement traditional cameras (RGB-imaging sensors) to perceive the environment more robustly. Spurred on by events, such as the 2004,

2005 Defence Advanced Research Projects Agency (DARPA) Grand Challenge and the 2007 DARPA Urban Challenge, the configuration of LiDAR and camera has become commonplace in almost all AV design [49], [50].

It is not just in AV research that LiDAR has been adopted as a data-gathering device. AAL research frequently uses LiDAR to make it easier for impaired people to navigate [51], [52]. Moreover, low-cost versions of the same LiDAR sensors are starting to enter the commercial world [53], [54], making them more accessible. This trend has led some to believe that LiDAR data, fused with image data, is the future of photography [55]. By removing the sensor's mechanical components, the industry hopes to reduce the cost dramatically [27].

Concerning ML methods used for activity recognition, CNNs appear to be a popular choice [56]–[58]. Regrettably, point cloud data is not a natural input to CNN. CNNs were designed to process RGB images or a suitable data format, therefore adapting them to work on 3-D data is not a straightforward extension. Depending on the sensor FoV, point cloud data is unstructured, unordered, prone to missing points, and affected by noise [7].

Early attempts at training CNNs using point cloud data required transforming the 3-D data to a series of RGB images at multiple views. These networks learned the depth map of the scene rather than the 3-D objects [59]. Although some success was attained, information was lost in the transformation, and the accuracy of the network suffered [60], [61].

More recently, a deep-learning (DL) architecture called PointNet [62], [63] was proposed. During training, PointNet takes unordered data points as a set of functions and maps a point set onto a vector. When making a classification, the vector representation is checked against the patterns identified in the dataset.

The advantage of PointNet is an end-to-end network that can process 3-D data without translating it to a complex representation [63]. While PointNet is good at making a classification, it is void of a mechanism to facilitate detection and localization. Moreover, the process is limited by available memory and lacks an understanding of the relationship between data points.

From the reviewed material, we understand the ML methods used for HAR vary between RNNs and CNNs, and one of the three different types of sensors is used to capture data. In terms of practicalities, we know that wearable sensors can produce some exemplary results but are not suitable considering the infrastructure requirements and the ease at which data can be influenced. While RGB or RGB-D sensors capture more information, they are susceptible to lighting conditions and, therefore, not reliable.

Therefore, a solution for classifying human activities needs to rely on data that the sensors are providing, use it intelligently to rain in dimensionality and adequately capture all the necessary attributes without the need for ubiquitous infrastructure. This approach would have to borrow from multimodal ML techniques, sensor data fusion, and given the diverse environments humans interact with, the algorithm needs to be robust, with relative ease.

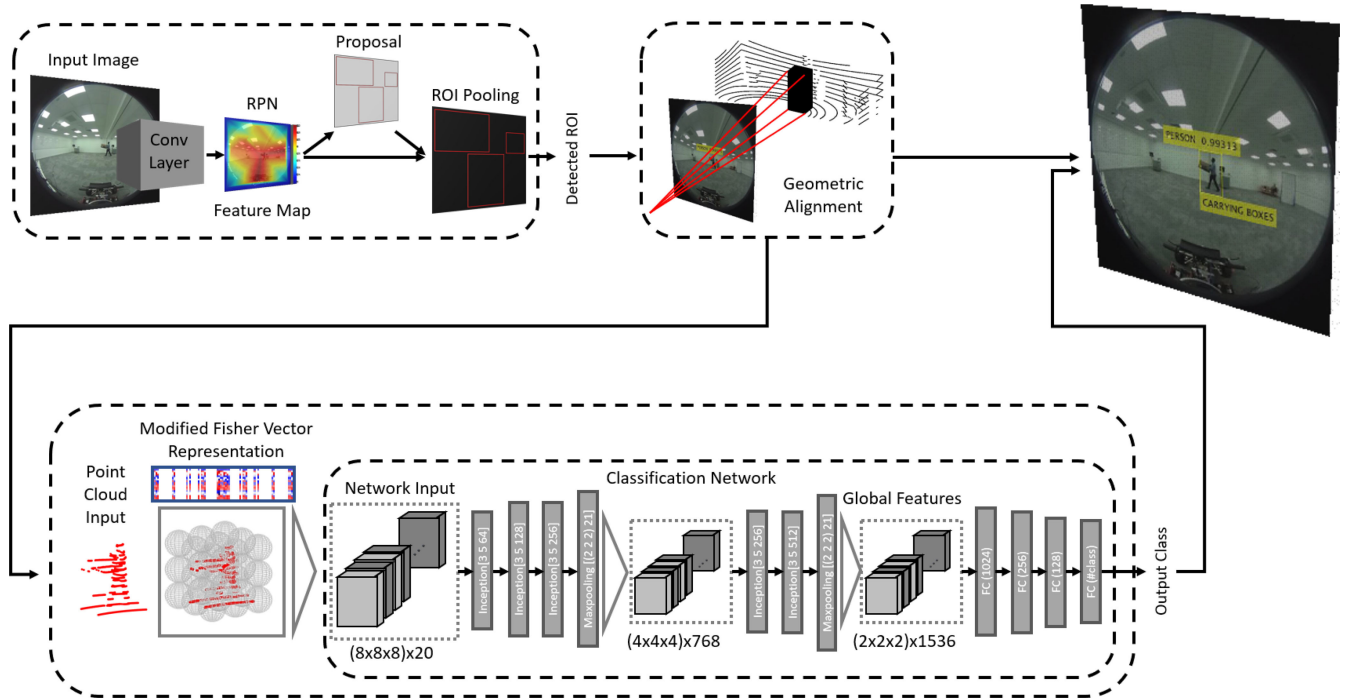


Fig. 2. Proposed MfV Net. We first leverage the power of a CNN to detect objects and propose an ROI. The geometrically aligned and translated point cloud data allows us to identify the corresponding 3-D ROI in the 3-D sample. This segmented ROI is converted into a modified Fisher vector representation before being passed onto the classification network. The resultant class is overlaid on the image, indicating the location and activity of the subject.

III. METHOD

In this section, we present our method of HAR using multimodal ML and sensing redundancy. The proposed method is based upon an advanced CNN architecture. The proposed framework illustrated in Fig. 2 utilizes a robust object detector to identify the ROI and a point cloud classifier to identify the subject's actions. The process requires the 2-D ROI to be translated and aligned to the LiDAR data before the segmented point cloud data is classified into the activity being performed.

A. Object Detection

The principal component of the proposed architecture detects and localizes agents before extracting the ROI from the RGB data. We modified ResNet [44] by training an RPN and a region classification network (RCN) to extract the ROIs. Referred to as a Faster-RCNN [6], this technique performs the convolution operation once per region before generating the feature map for the data. An extension of the RCNN and the Fast RCNN [6], the Faster RCNN's base classification network was ResNet-50. Trained on the ImageNet dataset [64] and finetuned using a subset (RGB images only) of the Loughborough London HAR (LboroLdnHAR) [3] dataset. This portion of the proposed framework was validated using an intersection over union (IoU) between the ground-truth bounding boxes and the predicted ROI. IoU detects the difference between the annotated bounding boxes of the RGB images and the proposed ROI. Using pairs of anchors with sizes {30, 19; 60, 38; 120, 76}, the ROI was labeled positive (object of interest present) when accuracy was greater than 0.65 and negative (object of interest not present) when less than 0.35.

In this case, the accuracy the classifier returns for a region behaves as a threshold for the detector. During training, the shortest side of all images was scaled to 246 pixels. Trained using stochastic gradient descent with a learning rate of 0.0001 and momentum of 0.9, the RCNN network used patch size of 16×16 pixels.

B. Geometric Alignment

The process of geometric alignment requires knowledge about proximity, orientation, and sensor modality. In terms of modality, there are vast differences between sensors that capture RGB, RGB-D, and point cloud data. For example, point cloud data is a representation of a collection of 3-D coordinates in space. Predominantly, these coordinates can be captured in two ways: 1) 3-D scanning and 2) photogrammetry [65]. The data are different in many ways to the structure and order of RGB data [61]. For example, the CCD or CMOS sensor at the heart of the camera captures reflected light from a subject. The specific location on the sensor that the reflected light strikes are recorded along with the true color value of the data there capturing. Conversely, point cloud data captures the location of an object in 3-D space. There is no sequence in which the ranges are recorded, and the scans vary in resolution, return either spherical, cylindrical, or Cartesian coordinates of the objects they are sensing.

We address these issues by translating and aligning different sensor modalities. This process is performed in two stages. The foremost stage translates and aligns RGB-D data captured by the Kinect sensor to the point cloud data captured by the LiDAR. The second stage translates the combined 3-D data streams to the RGB camera data. It should be noted that

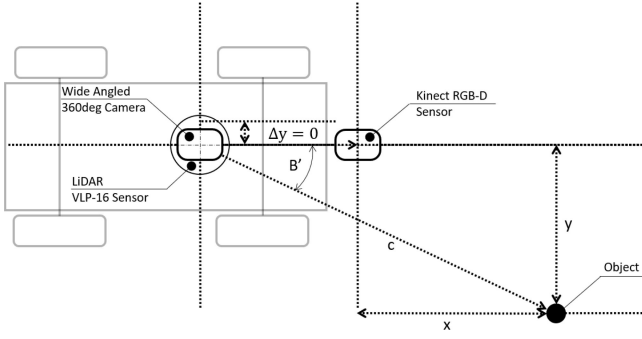


Fig. 3. Plan view of the setup showing the position of the sensors relative to the testbed.

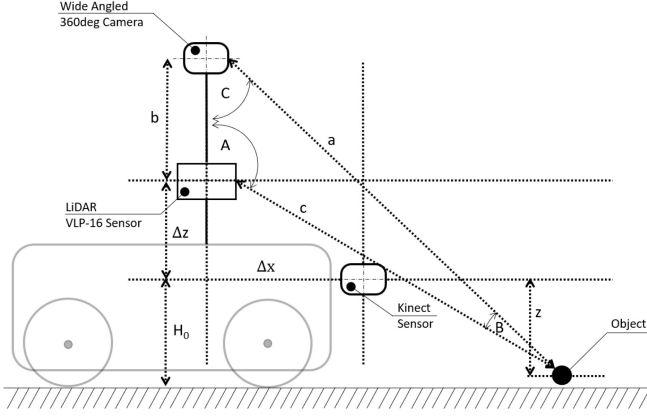


Fig. 4. Elevation view of the setup showing the position of the sensors relative to the testbed.

in the first stage, the data types are comparable, so the process is relatively simple. In the latter, the process is more complicated since we are translating different kinds of data. Figs. 3 and 4 illustrate a plan and elevated view of the sensor setup. This data is imperative to know so we can find the corresponding data points between different sensor modalities.

For the process of geometric alignment, consider a point in space represented by a 3-D coordinate. To shift that point's origin requires a scalar for each coordinate. For this explanation, consider an object O 4.76 m (x) in front of the Kinect sensor, 2.75 m (y) to the right, and 0.9 m (z) below the horizontal axis.

If we know that the Kinect sensor is offset from the LiDAR—0.45 m on the (Δx) axis, 0 m on the (Δy) axis, and 0.5 on the (Δz) axis—it is a matter of adding the scalar to corresponding coordinates, to align the two data streams as per the set of

$$X_T = \Delta x + x \quad (1)$$

$$Y_T = \Delta y + y \quad (2)$$

$$Z_T = \Delta z + z. \quad (3)$$

In the second stage of the process, we find the corresponding 3-D point for specific pixel output by the RGB sensor. This process is somewhat more complex than the foremost stage as the data types are not comparable.

For this derivation, consider an object O at a distance (c) of 5.6 m from the LiDAR. In this case, the object (O) is identified at an azimuth angle (B') of 3° and a zenith angle of 99° . The vertical distance (b) between the LiDAR and the RGB sensor

is 27 cm, and the RGB sensor is positioned ($H_0 + \Delta z + b$) 122 cm above the ground. Considering the distance c between the object O and the LiDAR sensor, we can describe B , the relative RGB azimuth angle as

$$B = \sin^{-1} \left(\frac{\sin(A) \times b}{a} \right) \Rightarrow \sin^{-1} \left(\frac{\sin(99^\circ) \times 0.27}{5.62} \right) = 2.7^\circ. \quad (4)$$

In this case, the RGB and LiDAR sensors are on the same longitudinal axis, and we know the LiDAR zenith angle A . As such, we can describe the RGB zenith angle C as

$$C = 180 - A - B \Rightarrow 180 - 99 - 2.72 = 78.2^\circ. \quad (5)$$

The distance a between object O and the RGB sensor can be described as

$$a = \frac{\sin(C) \times a}{\sin(A)} \Rightarrow \frac{\sin(78.28) \times 5.62}{\sin(99^\circ)} = 5.5 \text{ m}. \quad (6)$$

The purpose of this operation is to translate the LiDAR data stream to the image data. We assume that the longitudinal axis of the camera and the RGB sensor are aligned; however, an offset can be accounted for should it be required.

C. Point Cloud Classification

Unlike 2-D RGB data, 3-D data does not contain valuable descriptors—HOG, hue, saturation, and variance (HSV), etc.—used by feature learning networks to classify the objects of interest. By its nature, 3-D data needs to be modified before the CNN can classify it. The proposed method for classifying point cloud data consists of two main modules. The first module converts 3-D data to the modified Fisher vector representation. The second module processes the modified Fisher vector representation in a CNN before making a classification. An in-depth description of the point cloud classifier used in this research can be found in [7].

1) *Fisher Vector Representations*: Intraclass variability describes variations in appearance between two views of the same class or variations in appearance between two instances of the same class. In activity recognition, for example, significant variations in appearance between two instances can occur. Fisher vectors are a way of dealing with intraclass variability and determine the likelihood of getting specific data given an underlying theory.

A 2-D representation of the Fisher vector is shown in Fig. 5. Formally, it is the expected value of the observed information. This can best be described as having a probability distribution with a very sharp peak in the ideal circumstance. Conversely, we can obtain a distribution where the curve is very broad. This circumstance means there is a good deal of likelihood over a broad range of points.

Using a GMM, we can quantify the rate of curvature by looking at the probable distribution of the different points. It is often referred to as the curvature vector. It describes how curved the likelihood function is around the maximum. In simpler terms, the bigger the Fisher vector, the more curved the distribution is, meaning the more constrained the data is for that scenario.

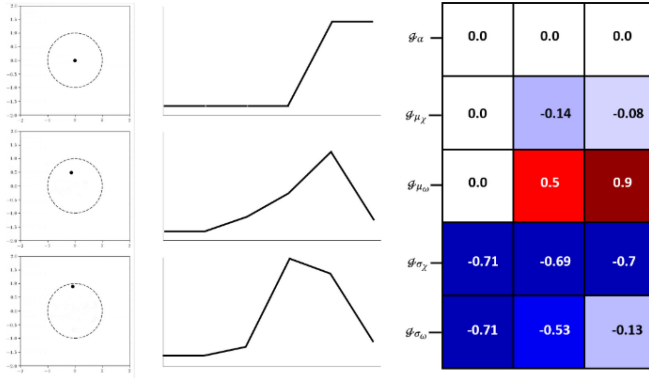


Fig. 5. Three 2-D points superimposed on a 2-D Gaussian on the left of the image. The center image is the FV, and the right image is the FV representation for the three points. The vector in the center of the image indicates the rate of curvature for the data.

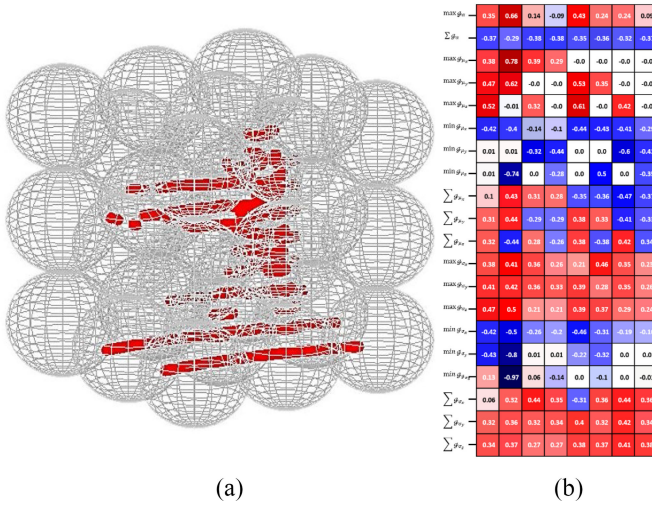


Fig. 6. (a) Spherical Gaussians superimposed on the point cloud data. (b) Modified Fisher vector representation of the GMM for the point cloud data.

Placing spherical Gaussians on a coarse grid provides structure, size, and the foundation for representing the image. If we compute each point derivative with respect to the Gaussian parameters and aggregate the results using three symmetric functions—maximum, minimum, and summation—we obtain a Fisher vector representation of a constant size that is invariant to permutations for different numbers of points.

For multiple points in a GMM, which consists of several Gaussian positioned on a grid, the Fisher vector representation can be interpreted as a statistically unique fingerprint of the 3-D data [66]. Fig. 6(a) depicts the spherical Gaussians superimposed on the point cloud data, and Fig. 6(b) shows the modified Fisher vector representation. It should be noted that the modified Fisher vector representation in Fig. 6(b) is for visualization purposes only.

2) *Classification of Fisher Vector Representation:* The main parts of the network used to classify the representation comprise several inception modules, max-pooling layers, and four fully connected layers [67], [68]. Training is done using backpropagation and standard SoftMax and cross-entropy loss

with batch normalization. Dropout is placed on the fully connected layers to prevent excessive co-adapting and overfitting during training. Dropout refers to a technique of ignoring specific neuron's response during a forward or backward pass after a certain time. Between the last max-pooling layer and the first fully connected layer, the network has approximately 4.6 million trained parameters.

The overall objective of the inception module is to overcome the dimensionality of the multiplier effect. An inception module allows us to use a number of filters in a single module [68]. While this will make the network architecture wider and more complex, the process works remarkably well in view of what we are trying to achieve.

The penultimate element of the classification network is a max-pooling layer. The max-pooling layer allows us to reduce the dimensionality of the inception module further.

Max-pooling layers work by passing all output elements of the inception module through an $n \times n$ filter. The $n \times n$ filter passes over the data with a fixed stride. Taking the first $n \times n$ region, we calculate the max value for that region before passing it on to the next. This process is repeated—passing the filter overall data elements—using the stride value to shift the pooling filter over in increments.

The final component in the classification network is four fully connected layers. The output of the previous layers is flattened into a single vector of values. These represent a probability that a particular feature belongs to a specific class. For example, in HAR, when a person is running, the fully connected layers should identify features representing the action with a high probability.

IV. EXPERIMENTAL RESULTS

This section is organized into six parts. The first section summarizes the LboroLdnHAR dataset developed in [3]. The second, third, and fourth sections report on the performance of the proposed framework, PointNet [62], and the deep neural-network framework (DNN) [3], [46], for HAR, respectively.

Comparative evaluation of the framework in this manner is essential to appraise the performance of the proposed process against state-of-the-art methods. It is not always clear how to do this, as what works for one framework will not necessarily work for another, particularly, when they process different data types. With this in mind, we have explained the metrics we use and the reason for using them.

In the fifth section, we benchmark the proposed framework against baseline methods of HAR using the SYSU-3D dataset before presenting some visible results of the proposed framework in the final section.

A. Dataset

Existing techniques for HAR are sensitive to viewpoint variations. This is a consequence of extracting features from viewpoint-dependent images. In contrast, we mine RGB data to identify a region in a point cloud scan. Since we segment and classify viewpoint-independent data, this technique is insensitive to viewpoint variations. To achieve this, we require a specific type of dataset to fulfill the requirements

of multimodality while providing optical and range data from at least two sensor streams of a known location.

The LboroLdnHAR [3] dataset consists of 6712 point cloud, RGB-D, and RGB annotated by hand samples. Data were captured by three sensors, a VLP-16 LiDAR, Kinect, and 360° camera. The dataset contains data of nine typical human activities performed by 16 participants. Activities were chosen on the basis of realistic activities performed in an office environment. The activities were carrying boxes, lying down, pushing a board, running; sitting on a chair; sitting on a stool; standing while texting; and walking and walking while texting.

The mean period for each data capture was approximately 35 s. Subjects started and ended the data capture periods with a T pose—standing upright with arms outstretched. During the experiments, the RGB 360° camera, RGB-D Kinect V2, and VLP-16 LiDAR sensor logged data at 30, 30, and 5 frames/s, respectively.

We split the dataset into three subsets. The first subsets consisted of images with annotated ground-truth ROI labels and trained the object detector. The second subsets contained aligned Kinect and LiDAR data. This subset of data was used to train the point cloud classifier. The final subset was a small number of the aligned and translated point cloud, RGB-D and RGB samples not used during the training or validation of the proposed framework. This section of the dataset was used to demonstrate the performance of the proposed framework visually.

B. Proposed Framework Performance

In this section, we report on the performance of the individual components of the proposed framework. We chose to evaluate the proposed framework in this manner to assist in the selection of a point cloud classifier. Not all point cloud classifiers are suitable for activity recognition. Some do not retain the crucial relationships between different points that allow us to understand the system of rigid segments connected by articulated joints that enable humans to move. Similarly, when comparing the performance of the proposed framework, it would make little sense to compare against a framework that used wearable MEMS as the data is dramatically different. To that end, we chose to evaluate the individual components of the framework separately and compare the methods of classification against an alternative point cloud classifier and a DNN rather than a classifier that processes MEMS data.

1) *Object Detection Performance*: The location and detection of agents were performed using an object detector. We trained and evaluated the object detector on the first subset of the LboroLdnHAR dataset [3]. To measure the detector's performance, we focused on the average precision—shown in Fig. 7—the log-average miss rates—shown in Fig. 8—and the F-score detailed in Table I. The average precision is a method of evaluation that incorporates the ability of the detector to make correct classifications, that is, precision, and the ability of the detector to find all relevant objects, that is, recall.

The log-average miss rate was attained by varying the thresholds on the detector confidence prediction and measure the rate of change in the true and false positives. Generally speaking, the log-average miss rate returns the results of the

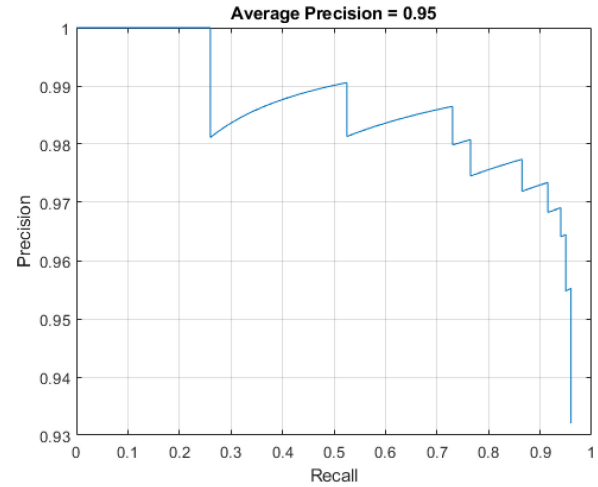


Fig. 7. Average precision of the object detector portion of the proposed framework showing the tradeoff between precision and recall.

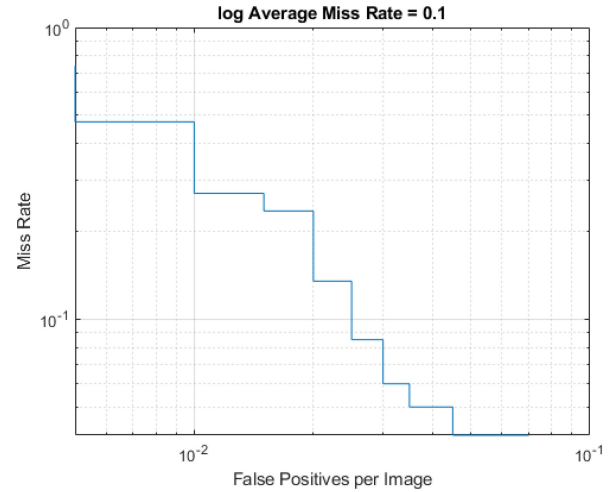


Fig. 8. Log-average miss rate of the object detector portion of the proposed framework indicating the quality of detection of the object detector.

TABLE I
OBJECT DETECTION PERFORMANCE

	PERFORMANCE
PRECISION	0.95
RECALL	0.96
LOG-AVERAGE MISS RATE	0.1
F-SCORE	0.95

object detector compared to the ground-truth table. This metric is another method of measuring the performance of the object detector [69], [70]. In this case, we found that the average precision and log-average miss rate for our detector were 0.95 and 0.1, respectively.

The F-score considers both the precision and the recall of the detector. Using binary classification in conjunction with the F-score allows us to determine the framework's accuracy [71]. Defined as two times the precision times the recall divided by the precision plus the recall, it reaches its best value at 1 and worst at 0. We determined the F-score—listed in Table I—as 0.95 for the object detector. These metrics are frequently used to determine the performance of an object detector.

TABLE II
PROPOSED FRAMEWORK PERFORMANCE TESTED ON A SUBSET OF THE LBOROLDN HAR DATASET

	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
PRECISION	1	1	0.988	0.726	0.881	0.968	0.9134	0.722	0.864
RECALL	0.959	1	1	0.688	0.890	0.957	0.989	0.776	0.795
F-SCORE	0.979	1	0.994	0.706	0.885	0.962	0.95	0.748	0.828
SAMPLE NO	74	103	86	73	101	94	104	72	81

TABLE III
POINTNET PERFORMANCE TESTED ON A SUBSET OF THE LBOROLDN HAR DATASET

	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
PRECISION	0.013	0.057	0.87	0	0.059	0	0	0	0
RECALL	0.019	0.714	0.11	0	0.038	0	NAN	NAN	NAN
F-SCORE	0.016	0.106	0.20	0	0.046	0	0	0	0
SAMPLE NO	74	103	86	73	101	94	104	72	81

Confusion Matrix

Output Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
CARRYING BOXES	71 9.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
LYING DOWN	0 0.0%	103 13.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
PUSHING A BOARD	0 0.0%	0 0.0%	86 10.9%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	98.9% 1.1%
RUNNING	0 0.0%	0 0.0%	0 0.0%	53 6.7%	2 0.3%	0 0.0%	8 1.0%	10 1.3%	72.6% 27.4%
SITTING ON A CHAIR	3 0.4%	0 0.0%	0 0.0%	5 0.6%	89 11.3%	3 0.4%	0 0.0%	1 0.1%	88.1% 11.9%
SITTING ON A STOOL	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1 0.1%	91 11.6%	1 0.1%	0 0.0%	96.8% 3.2%
STANDING WHILE TEXTING	0 0.0%	0 0.0%	0 0.0%	0 0.0%	8 1.0%	1 0.1%	95 12.1%	0 0.0%	91.3% 8.7%
WALKING	0 0.0%	0 0.0%	0 0.0%	14 1.8%	0 0.0%	0 0.0%	0 0.0%	52 6.6%	72.2% 27.8%
WALKING WHILE TEXTING	0 0.0%	0 0.0%	0 0.0%	5 0.6%	0 0.0%	0 0.0%	0 0.0%	6 0.8%	86.4% 13.6%
	95.9% 4.1%	100% 0.0%	100% 0.0%	68.8% 31.2%	89.0% 11.0%	95.8% 4.2%	99.0% 1.0%	77.6% 22.4%	79.5% 20.5%
								90.3% 9.7%	
Target Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING

Fig. 9. Confusion matrix for the classifier portion of the proposed framework.

2) *Point Cloud Classifier Performance*: Table II reports the average score for each class when tested with the LboroLdnHAR dataset [3]. To measure the classifier's performance, we focused on the precision, recall, and the F-score of each class. Fig. 9 shows the confusion matrix for the point cloud classifier. The y-axis is the output class, and the x-axis is the target class. The diagonal cells indicate true positives that are correctly classified. The off-diagonal cells indicate false positives that are incorrectly classified. The overall accuracy for the classifier was 0.903. The precision and recall were 0.895 and 0.894, respectively.

C. PointNet Performance

We compared the performance of the proposed method against PointNet [62]. PointNet [62] was trained and tested

Confusion Matrix

Output Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
CARRYING BOXES	1 0.1%	8 1.0%	0 0.0%	3 0.4%	30 3.8%	5 0.6%	0 0.0%	1 0.1%	3 0.4%
LYING DOWN	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
PUSHING A BOARD	0 0.0%	0 0.0%	5 0.6%	0 0.0%	2 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
RUNNING	67 8.5%	21 2.7%	77 9.8%	67 8.5%	35 4.5%	76 9.7%	96 12.2%	62 7.9%	84 10.7%
SITTING ON A CHAIR	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
SITTING ON A STOOL	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
STANDING WHILE TEXTING	0 0.0%	0 0.0%	2 0.3%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
WALKING	2 0.3%	73 9.3%	1 0.1%	7 0.9%	3 0.4%	14 1.8%	0 0.0%	4 0.5%	1 0.1%
WALKING WHILE TEXTING	3 0.4%	1 0.1%	2 0.3%	0 0.0%	30 3.8%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
	1.4% 98.6%	0.0% 100%	5.7% 94.3%	87.0% 13.0%	0.0% 100%	0.0% 100%	0.0% 100%	6.0% 94.0%	9.8% 90.2%
Target Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING

Fig. 10. Confusion matrix for PointNet. PointNet was tested on the same data used to test the proposed framework.

with the LboroLdnHAR dataset [3]. Table III reports the average score for each class from PointNet. Similar to the previous section, where we focused on the precision, recall, and F-score.

Fig. 10 shows the confusion matrix for PointNet [62]. The overall accuracy for the classifier was 0.098. The precision and recall were 0.111 and 0.147, respectively. Of particular interest was the network's inability to classify subjects lying down, sitting on a chair, and sitting on a stool. While the network was able to make some predictions, its performance was substandard compared to the 3-D classification network.

D. DNN Performance

To make an additional comparison, we benchmark the performance of the proposed method against a DNN framework for HAR reported in [3]. This is a similar framework reported in [46] and a frequently used method to identify

TABLE IV
DNN PERFORMANCE TESTED ON A SUBSET OF THE LBOROLDN HAR DATASET

	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
PRECISION	0.283	1	0.895	0.876	1	1	1	0.569	0.654
RECALL	0.677	1	0.987	0.771	0.980	1	0.990	0.577	0.441
F-SCORE	0.400	1	0.939	0.820	0.990	1	0.995	0.573	0.527
SAMPLE NO	74	103	86	73	101	94	104	72	81

Confusion Matrix

Output Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING
CARRYING BOXES	21 2.7%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	10 1.3%	67.7% 32.3%
LYING DOWN	0 0.0%	103 13.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
PUSHING A BOARD	0 0.0%	0 0.0%	77 9.8%	1 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	98.7% 1.3%
RUNNING	10 1.3%	0 0.0%	9 1.1%	64 8.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	77.1% 22.9%
SITTING ON A CHAIR	0 0.0%	0 0.0%	0 0.0%	2 0.3%	101 12.8%	0 0.0%	0 0.0%	0 0.0%	98.1% 1.9%
SITTING ON A STOOL	0 0.0%	0 0.0%	0 0.0%	0 0.0%	94 11.9%	0 0.0%	0 0.0%	0 0.0%	100% 0.0%
STANDING WHILE TEXTING	0 0.0%	0 0.0%	0 0.0%	1 0.1%	0 0.0%	104 13.2%	0 0.0%	0 0.0%	99.0% 1.0%
WALKING	7 0.9%	0 0.0%	0 0.0%	5 0.6%	0 0.0%	0 0.0%	41 5.2%	18 2.3%	57.7% 42.3%
WALKING WHILE TEXTING	36 4.6%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	31 3.9%	53 6.7%	44.2% 55.8%	83.5% 16.5%
	28.4% 71.6%	100% 0.0%	89.5% 10.5%	87.7% 12.3%	100% 0.0%	100% 0.0%	56.9% 43.1%	65.4% 34.6%	83.5% 16.5%
Target Class	CARRYING BOXES	LYING DOWN	PUSHING A BOARD	RUNNING	SITTING ON A CHAIR	SITTING ON A STOOL	STANDING WHILE TEXTING	WALKING	WALKING WHILE TEXTING

Fig. 11. Confusion matrix for the DNN framework. The DNN framework was tested on the same data used to test the proposed framework.

human activities from RGB-D data. The DNN framework was trained and tested with the LboroLdnHAR dataset [4]. Table IV reports on the precision, recall, and the F-score of each class from the DNN framework. Fig. 11 shows the confusion matrix for the DNN framework. The overall accuracy for the classifier was 0.835. The precision and recall were 0.808 and 0.825, respectively. Of particular interest was the network's performance in classifying subjects when moving. For example, the accuracy when classifying stationary subjects was remarkably higher than those from a dynamic class. This is evident in Fig. 11, where the framework frequently misclassifies walking, running, and carrying boxes.

E. Performance on the SYSU-3D Dataset

To benchmark the proposed framework, we trained and tested the network on the joint data from the SYSU-3D dataset. The SYSU-3D dataset consists of 40 subjects performing 12 activities: 1) drinking; 2) pouring; 3) calling phone; 4) playing phone; 5) wearing backpacks; 6) packing backpacks; 7) sitting chair; 8) moving chair; 9) taking our wallet; 10) taking from wallet; 11) mopping; and 12) sweeping. Fig. 12 shows the confusion matrix for the proposed framework when tested on the SYSU dataset. The overall accuracy for the classifier was 0.954. The precision and recall were 0.956 and 0.958, respectively. Table V reports

Confusion Matrix

Output Class	Calling Phone	Drinking	Mopping	Moving Chair	Packing Backpack	Playing Phone	Pouring	Sitting Chair	Sweeping	Taking From Wallet	Taking Our Wallet	Wearing Backpack
Calling Phone	1073 10.2%	3 0.0%	0 0.0%	0 0.0%	0 0.0%	98 0.9%	0 0.0%	0 0.0%	0 0.0%	2 0.0%	6 0.1%	8 0.1%
Drinking	4 0.0%	778 7.4%	0 0.0%	0 0.0%	1 0.0%	3 0.0%	0 0.0%	0 0.0%	0 0.0%	4 0.0%	0 0.0%	3 0.0%
Mopping	0 0.0%	0 0.0%	664 6.3%	1 0.0%	0 0.0%	0 0.0%	1 0.0%	11 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Moving Chair	0 0.0%	0 0.0%	0 0.0%	611 5.8%	3 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 0.0%
Packing Backpack	0 0.0%	0 0.0%	0 0.0%	2 0.0%	1171 11.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	3 0.0%
Playing Phone	11 0.1%	0 0.0%	0 0.0%	0 0.0%	0 0.0%	1131 10.8%	0 0.0%	0 0.0%	1 0.0%	6 0.1%	4 0.0%	0 0.0%
Pouring	1 0.0%	0 0.0%	0 0.0%	1 0.0%	3 0.0%	798 7.6%	0 0.0%	0 0.0%	0 0.0%	1 0.0%	0 0.0%	2 0.0%
Sitting Chair	2 0.0%	0 0.0%	0 0.0%	3 0.0%	1 0.0%	4 0.0%	461 4.4%	0 0.0%	0 0.0%	0 0.0%	14 0.1%	95.1% 4.9%
Sweeping	0 0.0%	0 0.0%	23 0.2%	1 0.0%	1 0.0%	0 0.0%	0 0.0%	648 6.2%	0 0.0%	0 0.0%	0 0.0%	0 0.0%
Taking From Wallet	7 0.1%	1 0.0%	0 0.0%	0 0.0%	1 0.0%	13 0.1%	0 0.0%	1 0.0%	1203 11.5%	126 1.2%	0 0.0%	4 0.0%
Taking Our Wallet	2 0.0%	1 0.0%	0 0.0%	0 0.0%	0 0.0%	7 0.1%	0 0.0%	4 0.0%	39 3.7%	570 5.4%	13 0.1%	89.5% 10.5%
Wearing Backpack	0 0.0%	0 0.0%	0 0.0%	4 0.0%	3 0.0%	3 0.0%	0 0.0%	1 0.0%	0 0.0%	0 0.0%	2 0.0%	896 8.6%
	97.5% 2.5%	99.4% 0.6%	96.7% 3.3%	97.9% 2.1%	99.1% 0.9%	89.6% 10.4%	100% 0.0%	98.3% 1.7%	98.2% 1.8%	95.9% 4.1%	80.3% 19.7%	94.7% 5.3%
Target Class	Calling Phone	Drinking	Mopping	Moving Chair	Packing Backpack	Playing Phone	Pouring	Sitting Chair	Sweeping	Taking From Wallet	Taking Our Wallet	Wearing Backpack

Fig. 12. Confusion matrix for the proposed framework trained and tested on the SYSU-3D dataset.

on the precision, recall, and the F-score of each class from the proposed framework. Table VI indicates a comparison of HAR techniques indicating that the proposed framework outperforms baseline methods—according to accuracy—by 8.5%.

Since the skeletal data provided in the SYSU-3D dataset can be regarded as noisy with large variations in viewpoint, the proposed framework brings significant performance improvement while showing how an activity recognition process can be insensitive to viewpoint variations.

F. Visual Results of the Proposed Framework

The purpose of the proposed algorithm is to assist an AV or AAL with HAR. Due to the multimodal requirements of the proposed framework, the results displayed here were a subset taken from the LboroLdnHAR dataset [3] while the experimental platform was stationary. All samples were gathered from indoor scenarios under controlled lighting.

The results presented in Fig. 13 Scenario 1 (c), Scenario 2 (c), Scenario 3 (c), and Scenario 4 (c) indicate that the proposed HAR framework performed as desired. In all cases, the subject performing the activity was detected, segmented, and classified correctly.

For example, in Fig. 13 Scenario 1 (a), the image shows a subject carrying a box. The first part of the proposed

TABLE V
PROPOSED FRAMEWORK PERFORMANCE TESTED ON A SUBSET OF THE SYSU-3D DATASET

	CALLING PHONE	DRINKING	MOPPING	MOVING CHAIR	PACKING BACKPACK	PLAYING PHONE	POURING	SITTING CHAIR	SWEEPING	TAKING FROM WALLET	TAKING OUR WALLET	WEARING BACKPACK
PRECISION	0.975	0.993	0.966	0.979	0.990	0.896	1	0.982	0.981	0.958	0.802	0.947
RECALL	0.901	0.978	0.980	0.988	0.995	0.980	0.988	0.950	0.962	0.887	0.894	0.985
F-SCORE	0.937	0.986	0.973	0.983	0.993	0.936	0.994	0.966	0.972	0.921	0.846	0.966
SAMPLE NO	1100	783	687	624	1182	1262	798	469	660	1255	710	946

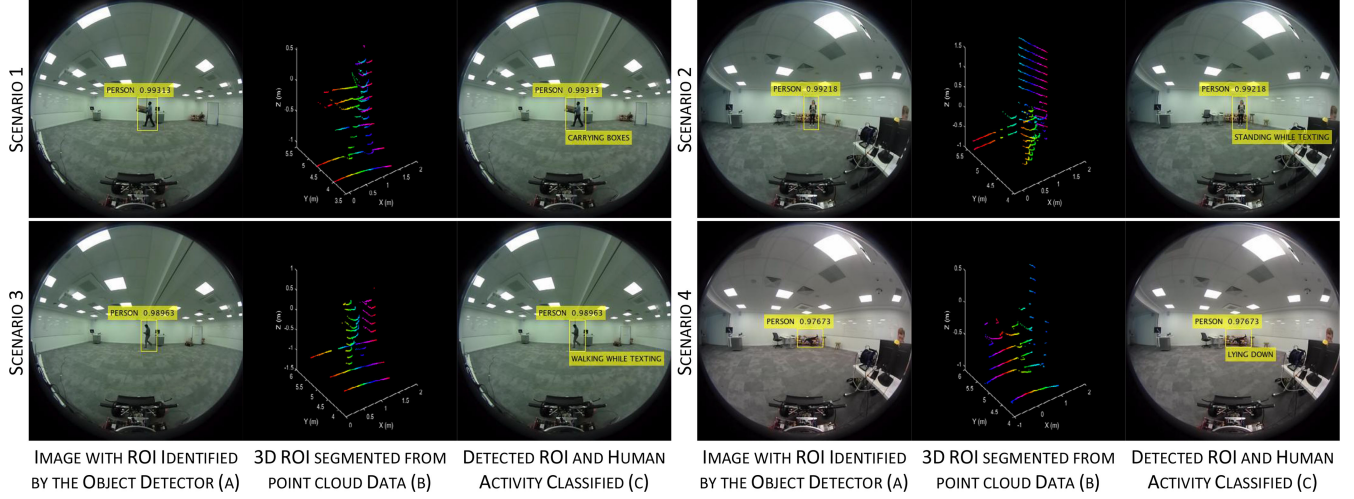


Fig. 13. Scenario 1 showing the subject carrying a box. Scenario 2 showing the subject standing while texting. Scenario 3 showing the subject walking while texting. Scenario 4 showing the subject lying down. Images in the first column indicate the output of the object detector. The center column shows the segmented point cloud data. The final column shows the output of the network and the associated class identified for the activity being performed. (a) Image with ROI identified by the object detector. (b) Three-dimensional ROI segmented from point cloud data. (c) Detected ROI and human activity classified.

framework—object detector—identifies, with a high degree of confidence (99%), the location and region that the person occupies.

Fig. 13 Scenario 1 (b) shows the 3-D ROI identified by the object detector in the 2-D image. Although there are unwanted artifacts located in the 3-D ROI, the activity being performed by the subject obtains correctly classified, as shown in Fig. 13 Scenario 1 (c).

In Fig. 13 Scenario 1 (a), there is a person—other than the subject—in the background. In this case, the object detector did not function as desired. This can be attributed to the design of the Faster-RCNN network, where input images with an aspect ratio of 360×360 are reduced down in size before processed.

It is important to remember, when processing images, the Faster-RCNN extracts features within the first few layers. When dealing with a small object in small images, the feature can, in effect, disappear in the middle of the network before detection occurs. In this case, the person in the background is not detected and, therefore, their activity is never classified in the latter part of the network. Regardless of the framework missing this subject, the results show how multimodal ML can be successfully applied to HAR when using RGB data captured by a camera and point cloud data captured using Kinect or LiDAR.

A similar but somewhat different issue occurs in Fig. 13 Scenario 2 (a). In this image, the subject is standing while texting. The object detector correctly identifies

TABLE VI
COMPARISON OF ACTION-RECOGNITION ACCURACY (%) WITH STATE-OF-THE-ART METHODS ON THE SYSU-3D DATASET

METHOD	YEAR	ACCURACY
ST-LSTM [72]	2018	73.4%
DYNAMIC SKELETONS [73]	2017	75.5%
DPRL [74]	2018	76.9%
VA-LSTM [75]	2017	77.5%
COMPLETE GR-GCN [76]	2019	77.9%
LOCAL+LGN [77]	2020	83.1%
ELEATT-GRU [78]	2020	85.7%
VA-FUSION [75]	2018	86.7%
SGNet [79]	2020	86.9%
PROPOSED FRAMEWORK	2020	95.4%

the ROI for the subject but misses the person behind the computer due to occlusion. Contrary to this, the subject of interest was detected correctly, and the corresponding 3-D ROI was extracted from the 3-D data. Of particular interest is Fig. 13 Scenario 2 (b). In this image, the subject is standing upright with hands out front, clasping a phone. This ROI is passed onto the classifier component of the proposed network before the correct result is overlaid onto the image shown in Fig. 13 Scenario 2 (c).

It is important to note that the phone in Fig. 13 Scenario 2 (b) occupies 19 3-D points. Conversely, the person sitting in the background in Fig. 13 Scenario 1 (a) occupies 79 pixels. Given that the object detector failed to identify the person in

the background, it infers that the point cloud classifier is adept at identifying point cloud features regardless of proximity.

A further example of a dynamic activity is presented in Fig. 13 Scenario 3 (a), where a subject is walking while texting. In this image, the subject was identified with 98% confidence. The corresponding 3-D ROI was segmented using the ROI identified by the object detector is shown in Fig. 13 Scenario 3 (b). As in Fig. 13 Scenario 2 (b), the subject in this image had their hands out front clasping a phone. This ROI was passed onto the classifier component before the correct result is overlaid onto the image shown in Fig. 13 Scenario 3 (c).

Fig. 13 Scenario 4 (a) shows a subject lying down. In this image, the area occupied by the subject, and some of the couch, was identified with 97% confidence. Fig. 13 Scenario 4 (b) shows the 3-D ROI identified by the object detector. In this image, the ROI contains the 3-D data of the subject and a portion of the couch. When passed onto the classifier, the correct result was overlaid onto the image. Much like Fig. 13 Scenario 1, where the subject was carrying a box, the object the subject is performing the activity with is incorporated into the classification processes.

V. DISCUSSION

In Fig. 13, we presented four scenarios displaying the different section of the proposed framework for HAR. In each case, the model outputs remarkably accurate results. Furthermore, we found that the proposed framework can make an accurate prediction from partial data with few points.

For example, in Fig. 13 Scenario 4 (b), the subject is in the supine position on a couch. The lack of point cloud data cannot be attributed to the instrument but the instrument principle of operation. Note the horizontal plains contacted by a laser—the further away, the greater the gap between plains.

In this case, points on the subject contacted by the laser are reduced to a single plain, making it difficult to classify. Moreover, given the obscurity and the absence of data points, humans trying to annotate the same data would find this a difficult task.

Notwithstanding the successes, we do observe a few intrinsic issues to heed. The primary and most pronounced issue is to do with the frame rate of the LiDAR (5 HZ). While possible to increase the frame rate, reducing the number of points captured by a single scan. Although the proposed framework shows resilience to frugal data captures, there is a point where the network will miss classify. Although possible to reduce the frame rate below 5 Hz and thus increase the data captured in a single frame. Doing so results in temporal ghosting of subjects moving at speed. When ghosting occurs, the RGB, point cloud and RGB-D data fall out of alignment. However, this can be managed with careful selection of the instruments frame rate.

We encountered other issues where the object detector failed to classify subjects at a distance from the camera. As commented in the results section, Fig. 13 Scenario 1 (a) showed a subject performing the task and a person sitting on the couch. In this scenario, the object detector misses the person on the

couch but identified the subject of interest. This issue was attributed to the distance the subject was from the RGB sensor. It is believed that this issue can be overcome by using sensors with a high-resolution image patch for far-away objects.

In terms of the point cloud classifier, it is important to note that the primary objective of all supervised learning framework is to discover the pattern linking the inputs X to the outputs Y , given a dataset D of size N [59], [60]. In the simplest terms, each input value has a dimension vector that represents the data we want the ML algorithm to learn. Commonly referred to as features, this representation is the understanding of X the algorithm develops during training. From the perspective of Y , the output can be anything, but the assumption is that it matches a categorical or nominal variable from the dataset used during training.

Generally, when using supervised learning methods for processing context data, a CNN obtains superior results with respect to traditional feature learning algorithms. Of course, in scenarios where data are sparse, a CNN can be outperformed. For example, when a CNN is trained on one dataset and expected to classify unfamiliar data that it has not been trained for, it will fail. However, when a CNN has been trained to recognize a specific action, such as falling, the framework should outperform all other forms of feature learning algorithms. Unfortunately, it is not always possible to acquire all the necessary data needed to train CNN to identify all specific actions. To that end, when we were collecting data, we focused on indoor activities where humans had a limited attention span. Had we included additional activities in the dataset used to train the proposed framework, it is clear from the results we presented in Section IV that they would have been classified with a high degree of accuracy.

In terms of comparison to other point cloud classifiers and their suitability for 3-D feature learning, we looked at [62], [80], and [81]. Researchers in [81] converted the point cloud data to a voxel grid array. How well the CNN acts on a voxel is dependent on the size of the grid—course or a fine. If applied to HAR, choosing a coarse grid leads to quantization and a substantial loss of information, and a fine grid increases the computational cost.

PointNet was another approach to classification that directly consumes unordered point clouds [62], [80]. This context refers to the fact that the model does not assume any spatial relationships between features. This is not the case for HAR, which assumes neighboring relations between the location of different points. Consequently, PointNet performs poorly when classifying HAR, regardless of how well it performs on classifying objects.

It should be noted that although the primary component of the proposed framework is a point cloud classifier. However, it is not the principal objective. The principal objective is to classify human activities robustly. With that in mind, we compared the proposed framework to a DNN framework for HAR [3], [46]. Although the DNN framework performed well, there are some intrinsic shortcomings. Primarily, the issues with the DNN framework are to do with the type of data processed. Sensors that capture RGB-D data are notoriously

susceptible to light changes [82] or interference from other sensors based upon structured light [83]. Consequently, the quality of light in the vision system environment is a key factor in its success. This is especially the case for networks that use sensors like the Kinect. Similarly, networks that process RGB data are affected to a lesser degree *D*.

VI. CONCLUSION

This article proposed a method of HAR using multimodal ML to process RGB, RGB-D, and LiDAR data. The algorithm consists of an object detector and a point cloud classification network. The object detector utilizes a Faster R-CNN to identify an ROI containing a person performing an activity. Translation and alignment of data types allow for the efficient segmentation of the point cloud data. The corresponding 3-D ROI is converted to a modified Fisher vector representation before a CNN classifies the activity. Performing HAR this way removes the need for wearable sensors to work in both indoor and outdoor environments. The technical challenges of the proposed framework lay with accurate alignment and translation of the different sensor modalities. Further technical challenges come in the form of sensor frequency of operation—which plays an integral role in perceiving the correct presence of subjects.

The object detection portion of the proposed framework accurately identifies the presence of subjects in an image, with an average precision of 95% and a log-average miss rate of 0.1, respectively. The classification portion of the proposed framework accurately classified the activities performed by the subject with 90.3%. The precision and recall were 89.5% and 89.4%, respectively. Compared to an alternative point cloud classifier, the proposed framework significantly outperformed the next suitable alternative, PointNet, by a staggering 80.5%. Furthermore, it was found that the proposed framework outperformed a state-of-the-art HAR framework using a DNN to process RGB-D by 6.9%.

To benchmark the proposed framework, we conducted a systematic evaluation using the SYSU-3D dataset. The proposed method in this article outperformed all reported baseline methods—according to accuracy—by 8.5%. Since the skeletal data provided in the SYSU-3D dataset can be regarded as noisy with large variations in viewpoint, the proposed framework brings significant performance improvement. It should be noted that the SYSU-3D dataset is substantially larger than the LboroLdnHAR dataset by approximately 100 thousand samples. This sample size difference can account for the varying performance of the proposed framework when trained on different datasets.

The implication of this research is machines that can better recognize human activities, without the need for wearable sensors, in varying environmental conditions and from various viewpoints. Indirectly, this research will allow machines to recognize when a person needs help, AVs to derive data-driven decision-making policies, and holds the potential to advance human-computer interaction to a new level of understanding.

REFERENCES

- [1] T. M. Bagautdinov, A. Alahi, F. Fleuret, P. Fua, and S. Savarese, "Social scene understanding: End-to-end multi-person action localization and collective activity recognition," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 3425–3434.
- [2] X. Liu and C.-S. Chua, "Multi-agent activity recognition using observation decomposed hidden Markov models," *Image Vis. Comput.*, vol. 24, no. 2, pp. 166–175, Feb. 2006.
- [3] M. Moencks, V. De-Silva, J. Roche, and A. Kondoz, "Adaptive feature processing for robust human activity recognition on a novel multi-modal dataset," 2019. [Online]. Available: arXiv:1901.02858.
- [4] R. Q. Mínguez, I. P. Alonso, D. Fernández-Llorca, and M. Á. Sotelo, "Pedestrian path, pose, and intention prediction through Gaussian process dynamical models and pedestrian activity recognition," *IEEE Trans. Intell. Transp. Syst.*, vol. 20, no. 5, pp. 1803–1814, May 2019.
- [5] V. De Silva, J. Roche, and A. Kondoz, "Robust fusion of LiDAR and wide-angle camera data for autonomous mobile robots," *Sensors*, vol. 18, no. 8, p. 2730, 2018.
- [6] S. Ren, K. He, R. Girshick, and J. Sun, "Faster R-CNN: Towards real-time object detection with region proposal networks," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 6, pp. 1137–1149, Jun. 2017.
- [7] Y. Ben-Shabat, M. Lindenbaum, and A. Fischer, "3DmFV: Three-dimensional point cloud classification in real-time using convolutional neural networks," *IEEE Robot. Autom. Lett.*, vol. 3, no. 4, pp. 3145–3152, Oct. 2018.
- [8] B. P. Yuhas, M. H. Goldstein, and T. J. Sejnowski, "Integration of acoustic and visual speech signals using neural networks," *IEEE Commun. Mag.*, vol. 27, no. 11, pp. 65–71, Nov. 1989.
- [9] X. Zhang *et al.*, "Emotion recognition from multimodal physiological signals using a regularized deep fusion of kernel machine," *IEEE Trans. Cybern.*, early access, May 14, 2020, doi: [10.1109/TCYB.2020.2987575](https://doi.org/10.1109/TCYB.2020.2987575).
- [10] A. Hong *et al.*, "A multimodal emotional human-robot interaction architecture for social robots engaged in bidirectional communication/vspace*7pt," *IEEE Trans. Cybern.*, early access, Mar. 6, 2020, doi: [10.1109/TCYB.2020.2974688](https://doi.org/10.1109/TCYB.2020.2974688).
- [11] Y. Yuan, Y. Lu, and Q. Wang, "Tracking as a whole: Multi-target tracking by modeling group behavior with sequential detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 12, pp. 3339–3349, Dec. 2017.
- [12] Y. Yuan, D. Wang, and Q. Wang, "Anomaly detection in traffic scenes via spatial-aware motion reconstruction," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 5, pp. 1198–1209, May 2017.
- [13] L.-P. Morency and T. Baltrušaitis, "Multimodal machine learning: Integrating language, vision and speech," in *Proc. 55th Annu. Meeting Assoc. Comput. Linguist. Tutorial Abstracts*, 2017, pp. 3–5.
- [14] T. Baltrušaitis, C. Ahuja, and L. P. Morency, "Multimodal machine learning: A survey and taxonomy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 2, pp. 423–443, Feb. 2019.
- [15] F. P. Such, V. Madhavan, E. Conti, J. Lehman, K. O. Stanley, and J. Clune, "Deep neuroevolution: Genetic algorithms are a competitive alternative for training deep neural networks for reinforcement learning," 2017. [Online]. Available: arXiv:1712.06567.
- [16] D. V. S. X. De Silva, W. A. C. Fernando, and S. T. Worrall, "Intra mode selection method for depth maps of 3D video based on rendering distortion modeling," *IEEE Trans. Consum. Electron.*, vol. 56, no. 4, pp. 2735–2740, Nov. 2010.
- [17] H. M. Le, P. Carr, Y. Yue, and P. Lucey, "Data-driven ghosting using deep imitation learning," in *Proc. MIT Sloan Sports Anal. Conf.*, Boston, MA, USA, Mar. 2017, pp. 1–12. [Online]. Available: <https://resolver.caltech.edu/CaltechAUTHORS:20170316-121646643>
- [18] V. Ghasemi and A. A. Pouyan, "Human activity recognition in ambient assisted living environments using a convex optimization problem," in *Proc. 2nd Int. Conf. Signal Process. Intell. Syst.*, 2017, pp. 1–6.
- [19] H. T. Malazi and M. Davari, "Combining emerging patterns with random forest for complex activity recognition in smart homes," *Appl. Intell.*, vol. 48, no. 2, pp. 315–330, 2018.
- [20] O. Banos, M. Damas, H. Pomares, F. Rojas, B. Delgado-Marquez, and O. Valenzuela, "Human activity recognition based on a sensor weighting hierarchical classifier," *Soft Comput.*, vol. 17, no. 2, pp. 333–343, 2013.
- [21] K. Gadzicki, R. Khamsehashari, and C. Zetsche, "Early vs late fusion in multimodal convolutional neural networks," in *Proc. 23rd Int. Conf. Inf. Fusion*, 2020, pp. 1–6.
- [22] G. Chetty, M. White, M. Singh, and A. Mishra, "Multi modal activity recognition based on automatic feature discovery," in *Proc. Int. Conf. Comput. Sustain. Global Dev. (INDIACom)*, 2014, pp. 632–637, doi: [10.1109/IndiaCom.2014.6828039](https://doi.org/10.1109/IndiaCom.2014.6828039).
- [23] F. Diederichs, T. Schuttke, and D. Spath, "Driver intention algorithm for pedestrian protection and automated emergency braking systems," in *Proc. IEEE Conf. Intell. Transp. Syst.*, 2015, pp. 1049–1054.

- [24] E. Rehder, H. Kloeden, and C. Stiller, "Head detection and orientation estimation for pedestrian safety," in *Proc. 17th IEEE Int. Conf. Intell. Transp. Syst.*, 2014, pp. 2292–2297.
- [25] P. Gupta and T. Dallas, "Feature selection and activity recognition system using a single triaxial accelerometer," *IEEE Trans. Biomed. Eng.*, vol. 61, no. 6, pp. 1780–1786, Jun. 2014.
- [26] E. Garcia-Ceja, R. F. Brena, J. C. Carrasco-Jimenez, and L. Garrido, "Long-term activity recognition from wristwatch accelerometer data," *Sensors*, vol. 14, no. 2, pp. 22500–22524, 2014.
- [27] Z. Chen, Q. Zhu, Y. C. Soh, and L. Zhang, "Robust human activity recognition using smartphone sensors via CT-PCA and online SVM," *IEEE Trans. Ind. Informat.*, vol. 13, no. 6, pp. 3070–3080, Dec. 2017.
- [28] Z. Zhang and S. Poslad, "Improved use of foot force sensors and mobile phone GPS for mobility activity recognition," *IEEE Sens. J.*, vol. 14, no. 12, pp. 4340–4347, Dec. 2014.
- [29] Y. Kwon, K. Kang, and C. Bae, "Unsupervised learning for human activity recognition using smartphone sensors," *Expert Syst. Appl.*, vol. 41, no. 14, pp. 6067–6074, 2014.
- [30] M. Janidarmian, A. R. Fekr, K. Radecka, and Z. Zilic, "A comprehensive analysis on wearable acceleration sensors in human activity recognition," *Sensors*, vol. 17, no. 3, p. 529, 2017.
- [31] R. Igual, C. Medrano, and I. Plaza, "A comparison of public datasets for acceleration-based fall detection," *Med. Eng. Phys.*, vol. 37, no. 9, Sep. 2015, pp. 870–878.
- [32] A. Stisen *et al.*, "Smart devices are different: Assessing and mitigating mobile sensing heterogeneities for activity recognition," in *Proc. 13th ACM Conf. Embedded Netw. Sens. Syst.*, 2015, pp. 127–140.
- [33] F. Attal, S. Mohammed, M. Dedabrishvili, F. Chamroukhi, L. Oukhellou, and Y. Amirat, "Physical human activity recognition using wearable sensors," *Sensors*, vol. 15, no. 12, pp. 31314–31338, 2015.
- [34] D. M. Karantonis, M. R. Narayanan, M. Mathie, N. H. Lovell, and B. G. Celler, "Implementation of a real-time human movement classifier using a triaxial accelerometer for ambulatory monitoring," *IEEE Trans. Inf. Technol. Biomed.*, vol. 10, no. 1, pp. 156–167, Jan. 2006.
- [35] L. Atallah, B. P. L. Lo, R. C. King, and G.-Z. Yang, "Sensor placement for activity detection using wearable accelerometers," in *Proc. Int. Conf. Body Sens. Netw.*, 2010, pp. 24–29.
- [36] J.-Y. Yang, J.-S. Wang, and Y.-P. Chen, "Using acceleration measurements for activity recognition: An effective learning algorithm for constructing neural classifiers," *Pattern Recognit. Lett.*, vol. 29, no. 16, pp. 2213–2220, 2008.
- [37] M. M. Hamdi, M. I. Awad, M. M. Abdelhameed, and F. A. Tolbah, "Lower limb gait activity recognition using Inertial Measurement Units for rehabilitation robotics," in *Proc. 17th Int. Conf. Adv. Robot.*, 2015, pp. 316–322.
- [38] X. Hu, C. Yao, and G. S. Soh, "Performance evaluation of lower limb ambulatory measurement using reduced inertial measurement units and 3R gait model," in *Proc. IEEE Int. Conf. Rehabil. Robot.*, 2015, pp. 549–554.
- [39] S. I. Patel and M. J. Souter, "Equipment-related electrocardiographic artifacts: Causes, characteristics, consequences, and correction," *Anesthesiology*, vol. 108, no. 1, pp. 138–148, 2008.
- [40] A. Vaidya, R. Bentley-Lewis, X. Jeunemaitre, G. K. Adler, and J. S. Williams, "Dietary sodium alters the prevalence of electrocardiogram determined left ventricular hypertrophy in hypertension," *Amer. J. Hypertens.*, vol. 22, no. 6, pp. 669–673, 2009.
- [41] D. B. Calne, J. Brennan, A. S. D. Spiers, and G. M. Stern, "Hypotension caused by L-dopa," *Brit. Med. J.*, vol. 1, no. 5694, pp. 474–475, 1970.
- [42] S. U. Park, J. H. Park, M. A. Al-Masni, M. A. Al-Antari, M. Z. Uddin, and T. S. Kim, "A depth camera-based human activity recognition via deep learning recurrent neural network for health and social care services," *Procedia Comput. Sci.*, vol. 100, pp. 78–84, Oct. 2016.
- [43] K.-T. Song and W.-J. Chen, "Human activity recognition using a mobile camera," in *Proc. 8th Int. Conf. Ubiquitous Robots Ambient Intell. (URAI)*, 2011, pp. 3–8.
- [44] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 770–778, doi: [10.1109/CVPR.2016.90](https://doi.org/10.1109/CVPR.2016.90).
- [45] S. Mukherjee, L. Anvitha, and T. M. Lahari, "Human activity recognition in RGB-D videos by dynamic images," *Multimed. Tools Appl.*, vol. 79, pp. 19787–19801, Mar. 2020.
- [46] E. Cippitelli, S. Gasparrini, E. Gambi, and S. Spinsante, "A human activity recognition system using Skeleton data from RGBD sensors," *Comput. Intell. Neurosci.*, vol. 2016, Mar. 2016, Art. no. 4351435.
- [47] H. Feng, S. Wang, H. Xu, and S. S. Ge, "Object activity scene description, construction, and recognition," *IEEE Trans. Cybern.*, early access, Apr. 1, 2019, doi: [10.1109/TCYB.2019.2904901](https://doi.org/10.1109/TCYB.2019.2904901).
- [48] M. M. Arzani, M. Fathy, A. A. Azirani, and E. Adeli, "Switching structured prediction for simple and complex human activity recognition," *IEEE Trans. Cybern.*, early access, Jan. 14, 2020, doi: [10.1109/TCYB.2019.2960481](https://doi.org/10.1109/TCYB.2019.2960481).
- [49] J. Byun, K.-I. Na, M. Noh, and S. Kim, "ESTRO: Design and development of intelligent autonomous vehicle for shuttle service in the ETRI," in *Proc. Workshop Plan. Percept. Navig. Intell. Veh.*, Portugal, 2012.
- [50] N. Suganuma and T. Uozumi, "Development of an autonomous vehicle—System overview of testride vehicle in the Tokyo motor show 2011," in *Proc. SICE Annu. Conf. (SICE)*, 2012, pp. 215–218.
- [51] A. Mancini, E. Frontoni, and P. Zingaretti, "Embedded multisensor system for safe point-to-point navigation of impaired users," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3543–3555, Dec. 2015.
- [52] T. Pallejà, M. Tresanchez, M. Teixidó, and J. Palacin, "Bioinspired electronic white cane implementation based on a LiDAR, a tri-axial accelerometer and a tactile belt," *Sensors*, vol. 10, no. 12, pp. 11322–11339, 2010.
- [53] *LiDAR—Overview of Technology, Applications, Market Features & Industry*, Centre for Applied Remote Sensing, Modelling and Simulation, Univ. Victoria, Victoria, BC, Canada, 2006.
- [54] R. Saracco, *LiDAR Hits Mass Market*, IEEE Future Directions, Piscataway, NJ, USA, 2020.
- [55] R. Saracco, *Light Field Technology—The Future of Photography?* IEEE Future Directions, Piscataway, NJ, USA, 2020.
- [56] F. Luo, S. Poslad, and E. Bodanese, "Temporal convolutional networks for multi-person activity recognition using a 2D LiDAR," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7432–7442, Aug. 2020.
- [57] J. Wang, Y. Chen, S. Hao, X. Peng, and L. Hu, "Deep learning for sensor-based activity recognition: A survey," *Pattern Recognit. Lett.*, vol. 119, pp. 3–11, Mar. 2019.
- [58] H. F. Nweke, Y. W. Teh, M. A. Al-garadi, and U. R. Alo, "Deep learning algorithms for human activity recognition using mobile and wearable sensor networks: State of the art and research challenges," *Expert Syst. Appl.*, vol. 105, pp. 233–261, Sep. 2018.
- [59] C. R. Qi, H. Su, M. Nießner, A. Dai, M. Yan, and L. J. Guibas, "Volumetric and multi-view CNNs for object classification on 3D data," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2016, pp. 5648–5656, doi: [10.1109/CVPR.2016.609](https://doi.org/10.1109/CVPR.2016.609).
- [60] Z. Wang, W. Zhan, and M. Tomizuka, "Fusing bird's eye view LiDAR point cloud and front view camera image for 3D object detection," in *Proc. IEEE Intell. Veh. Symp.*, 2018, pp. 1–6.
- [61] Y. Cai, T. Zhang, H. Wang, Y. Li, Q. Liu, and X. Chen, "3D vehicle detection based on LiDAR and camera fusion," *Automot. Innov.*, vol. 2, pp. 276–283, Nov. 2019.
- [62] C. R. Qi, L. Yi, H. Su, and L. J. Guibas, "PointNet++: Deep hierarchical feature learning on point sets in a metric space," in *Advances in Neural Information Processing Systems*. Red Hook, NY, USA: Curran Assoc., 2017.
- [63] C. R. Qi, H. Su, K. Mo, and L. J. Guibas, "PointNet: Deep learning on point sets for 3D classification and segmentation," in *Proc. 30th IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2017, pp. 77–85.
- [64] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet," in *Advances in Neural Information Processing Systems* 25. Red Hook, NY, USA: Curran Assoc., 2012.
- [65] D. W. James, J. Eckermann, B. Fawsi, and J. Sienz, "Point cloud data from Photogrammetry techniques to generate 3D Geometry," in *Proc. Conf. Assoc. Comput. Mech. Eng.*, 2015, p. 4.
- [66] R. A. Fisher, "Inverse probability," *Math. Proc. Cambridge Philos. Soc.*, vol. 26, no. 4, pp. 528–535, 1930.
- [67] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, 2015, pp. 1–9, doi: [10.1109/CVPR.2015.7298594](https://doi.org/10.1109/CVPR.2015.7298594).
- [68] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
- [69] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *Int. J. Comput. Vis.*, vol. 88, pp. 303–338, 2010.
- [70] S. Ceri *et al.*, *An Introduction to Information Retrieval*. Cambridge, U.K.: Cambridge Univ. Press, 2013.
- [71] P. Bhatia, *Data Mining and Data Warehousing: Principles and Practical Techniques*. Cambridge, U.K.: Cambridge Univ. Press, 2019.
- [72] J. Liu, A. Shahroudy, D. Xu, A. C. Kot, and G. Wang, "Skeleton-based action recognition using spatio-temporal LSTM network with trust gates," 2017. [Online]. Available: [arXiv:1706.08276](https://arxiv.org/abs/1706.08276).

- [73] J.-F. Hu, W.-S. Zheng, J. Lai, and J. Zhang, "Jointly learning heterogeneous features for RGB-D activity recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 39, no. 11, pp. 2186–2200, Nov. 2017.
- [74] Y. Tang, Y. Tian, J. Lu, P. Li, and J. Zhou, "Deep progressive reinforcement learning for skeleton-based action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 5323–5332.
- [75] P. Zhang, C. Lan, J. Xing, W. Zeng, J. Xue, and N. Zheng, "View adaptive recurrent neural networks for high performance human action recognition from skeleton data," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2017, pp. 2136–2145.
- [76] X. Gao, W. Hu, J. Tang, J. Liu, and Z. Guo, "Optimized skeleton-based action recognition via sparsified graph regression," in *Proc. 27th ACM Int. Conf. Multimedia*, 2019, pp. 601–610.
- [77] Q. Ke, M. Bennamoun, H. Rahmani, S. An, F. Sohel, and F. Boussaid, "Learning latent global network for skeleton-based action prediction," *IEEE Trans. Image Process.*, vol. 29, pp. 959–970, 2020.
- [78] P. Zhang, J. Xue, C. Lan, W. Zeng, Z. Gao, and N. Zheng, "EleAtt-RNN: Adding attentiveness to neurons in recurrent neural networks," *IEEE Trans. Image Process.*, vol. 29, pp. 1061–1073, 2020.
- [79] P. Zhang, C. Lan, W. Zeng, J. Xing, J. Xue, and N. Zheng, "Semantics-guided neural networks for efficient skeleton-based human action recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 1109–1118.
- [80] C. R. Qi, W. Liu, C. Wu, H. Su, and L. J. Guibas, "Frustum PointNets for 3D object detection from RGB-D data," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 918–927.
- [81] Y. Zhou and O. Tuzel, "VoxelNet: End-to-end learning for point cloud based 3D object detection," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, 2018, pp. 4490–4499.
- [82] J. Jancsek, D. Aleinikava, and G. M. Mirsky, "Optimizing kinect depth sensing using dynamic polarization," in *Proc. Conf. Integr. Technol. Comput. Sci. Educ. (ITICSE)*, 2017, pp. 767–768.
- [83] R. M. Martin, M. Lorbach, and O. Brock, "Deterioration of depth measurements due to interference of multiple RGB-D sensors," in *Proc. IEEE Int. Conf. Intell. Robots Syst.*, 2014, pp. 4205–4212.