IEEE Access
Multidisciplinary ⁝ Rapid Review ⁝ Open Access Journal

Received 7 November 2023, accepted 16 December 2023, date of publication 25 December 2023, date of current version 4 January 2024.

Digital Object Identifier 10.1109/ACCESS.2023.3347418

RESEARCH ARTICLE

Learning Spatial Affordances From 3D Point Clouds for Mapping Unseen Human Actions in Indoor Environments

LASITHA PIYATHILAKA[1], (Member, IEEE), SARATH KODAGODA[2], (Member, IEEE), KARTHICK THIYAGARAJAN[2], (Senior Member, IEEE), MASSIMO PICCARDI[3], (Senior Member, IEEE), D. M. G. PREETHICHANDRA[1], (Senior Member, IEEE), AND UMER IZHAR[4], (Member, IEEE)

[1]School of Engineering and Technology, Central Queensland University, Rockhampton, QLD 4700, Australia
[2]Faculty of Engineering and Information Technology, UTS Robotics Institute, University of Technology Sydney, Sydney, NSW 2007, Australia
[3]Faculty of Engineering and Information Technology, School of Electrical and Data Engineering, University of Technology Sydney, Sydney, NSW 2007, Australia
[4]School of Science Technology and Engineering, University of the Sunshine Coast, Sunshine Coast, QLD 4556, Australia

Corresponding author: Lasitha Piyathilaka (l.piyathilaka@cqu.edu.au)

ABSTRACT Many indoor robots operate in environments designed to support human activities. Understanding probable human actions in such surroundings is crucial for facilitating better human-robot interactions. This article presents an innovative approach to map unseen human actions in indoor environments by leveraging spatial affordances learned from geometric features extracted from point clouds captured by 3D cameras. Instead of directly observing real people to understand human context, the method utilizes virtual human models and their interactions with the environment to uncover hidden human affordances. This approach proves to be efficient for learning the affordance map, even when dealing with highly imbalanced datasets. To achieve this, we employ a supervised learning model optimized for the F1 score, using the Structured-SVM (S-SVM) architecture. We conducted experiments with actual 3D scenes, evaluating various affordance types both qualitatively and quantitatively. The results show that the proposed S-SVM-based method outperforms other models, demonstrating its effectiveness in efficiently mapping human context in indoor environments. The S-SVM-based method outperformed other models, demonstrating efficient human context mapping in indoor environments.

INDEX TERMS Human activity recognition, human centered robotics, human–robot interaction, smart sensing, spatial affordances.

I. INTRODUCTION

Intelligent sensing technologies play a significant role in reasoning human activities for challenging tasks in Human-Robot Interaction (HRI) [1]. As opposed to traditional robotics, where robots work in isolation, modern-day robotic applications require robots to interact safely in a way that is acceptable to humans. Therefore, it is believed that a robot's ability to understand the human context in its surroundings is the most vital cognitive ability that robots should possess for effective human-robot interaction.

In previous research, understanding the human context for robotic applications was accomplished by analysing human movement patterns [2], recognising human activities [3], [4], and simulating interactions between people and things in their surroundings [5]. In order to learn human context, almost all of these techniques require robots to recognise and track people for a considerable amount of time which often fails in congested and chaotic environments.

Affordance theory, as proposed by Gibson [6], posits that human actions are influenced by the arrangement of

The associate editor coordinating the review of this manuscript and approving it for publication was Jamshed Iqbal.

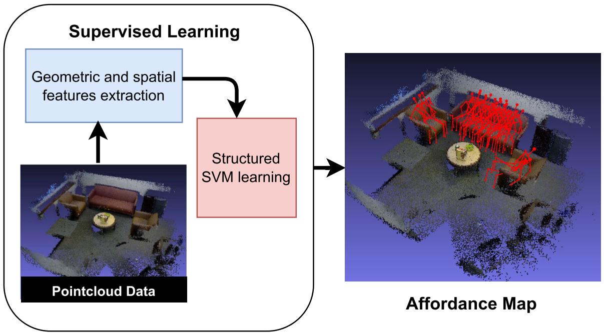


FIGURE 1. The spatial affordance mapping concept.

their environment. This close relationship between human context and environmental affordances can potentially be utilized to comprehend human context even in the absence of direct observation. The underlying idea is to derive environmental affordances solely from the geometric features of the environment, leading to the introduction of the concept called "affordance map" in this article. The affordance map involves the mapping of potential human actions in 3D scenes through virtual human models by using a supervised learning approach as shown in Fig. 1. This novel approach allows for the learning of human context within a given environment without the need to observe real humans directly. Instead, it infers virtual 3D human skeletons associated with each affordance type. By doing so, it circumvents the challenges associated with human detection and tracking, making it a promising method for understanding human behavior in various settings. Similar to how robots currently use grid-based maps for localization, route planning, and obstacle detection, incorporating an affordance map could significantly enhance human-robot interaction. For instance, a domestic service robot could leverage the human context data within the affordance map to organize objects in a configuration preferred by humans before they return home from work. This proactive arrangement would improve the robot's ability to assist and cater to human preferences. Additionally, the robot could utilize pose information from digital human models stored in the affordance map to efficiently find and locate various objects within the environment. This capability would enable the robot to perform tasks more accurately and swiftly, enhancing its overall utility in domestic settings.

The affordance map could serve as a valuable resource for predicting and detecting human activity, even when direct observation of people is not possible. By relying on the reliable priors provided by the affordance map, the robot can better understand and anticipate human behaviours, leading to improved adaptability and responsiveness during interactions with humans. In this article, we proposed a framework for learning spatial affordances using 3D point cloud-enabled geometric data in order toÂ build an affordance map in indoor environments such as a large room with multiple types of affordances. This work extends our previous work [7] based on cost-sensitive SVM, with the Structured SVM (S-SVM) framework which reported better overall results. The major contributions of our work are as follows:

1) Proposed a method for building affordance maps that examines the geometrical features of the surroundings to predict probable affordances as human skeleton models with affordance likelihoods.
2) Proposed an approach for efficiently learning affordance map with performance metric F1-score optimised Structured SVM (S-SVM) algorithm.
3) By using grid locations of the 3D point clouds from the test dataset, performedÂ a qualitative analysis of affordance mapping by plotting the skeleton model of detected affordances. The results reveal the proper alignment of each skeleton associated with each affordance type. The S-SVM based learner produced acceptable results even with the dataset's high-class imbalances.
4) In addition to qualitative assessment, all affordance categories explored in this work were quantitatively evaluated. The proposed S-SVM approach outperformed other relevant SVM learners.

The remainder of this article is structured as follows: Section II reviews related work, while Section III describes affordance detection using a supervised learning approach. Section IV explains a method for learning the affordance map using a structured SVM algorithm, and Section V proposes an approach for affordance learning using the performance metric F1-score optimised structured SVM. Section VI presents the results of quantitative and quantitative analysis, and Section VII concludes this article by summarising the key outcomes and briefly outlining the prospects.

II. RELATED WORK

The field of automatic recognition of human context within an environment has been extensively explored, with a substantial body of prior research [8], [9], [10], [11]. Much of this earlier research hinges on the necessity of detecting humans and identifying their activities as prerequisites for effectively understanding the human context within an environment. Consequently, a predominant focus among researchers has been on addressing challenges related to human tracking [9], [10], [11] and human activity detection [12], [13].

Several researchers have explored the use of robotic systems for human activity recognition as a means to learn human context [14], [15]. The primary advantage of this approach lies in the possibility of employing active sensing. However, learning human context through direct observation of real humans can be challenging in various scenarios. Firstly, robots often need to observe humans for a significant period before successfully learning the human context in a new environment. This extended observation time can be impractical and time-consuming. Secondly, many robots are required to operate in environments where humans are not present. For instance, a service robot might need to arrange items in a house before humans return home from work. In such situations, the robot must gather human context information without the opportunity to observe real humans directly. Thirdly, even when humans are visible, learning human context from them can still be difficult in certain scenarios. Challenges arise when the human subject moves out of the sensory range or when sensor inputs become obscured, hindering the robot's ability to effectively learn and comprehend human context.

To address these challenges, alternative approaches, such as utilizing affordance maps and virtual human models, can be employed. These methods allow robots to efficiently learn human context without solely relying on direct observation of real humans. By incorporating such techniques, robots can improve their interactions with humans and effectively assist them in various environments.

Several researchers have explored the possibility of learning hidden action affordances within environments by solely analyzing their geometric properties. These novel approaches aim to bypass the requirement of human detection and tracking. For instance, in the study [16], the researchers employed virtual human models to identify objects with a specific affordance, such as being *sittable*.To accomplish this, they utilized a human model in a sitting pose to calculate distance features, enabling the learning of an affordance model represented as a probability density. The affordance probabilities were then computed for each 3D grid location within the room, and the locations with the highest probabilities were classified as *sittable* areas.

While the researchers achieved promising recognition accuracies with synthetic datasets, their approach encountered challenges when tested in a real 3D environment. The recorded results were not as successful as expected, indicating that the model's performance did not generalize well beyond synthetic settings.

In the research conducted by [17], virtual human models were introduced to learn human-object relationships, also known as affordances, in 3D scenes. Initially, object-human relationships were learned from labelled objects. These learned models were then employed to arrange objects in a manner preferred by humans in a synthetic environment. However, a limitation of their approach was that it assumed objects were already detected and labelled, making it impractical for use in a new 3D room without object labels.

To mitigate this constraint, an extension of their research was introduced in [5]. In this extension, researchers proposed the utilization of an Infinite Factored Topic Model (IFTM) to capture object-human relationships and enhance the accuracy of object detection. They incorporated hallucinated humans to extract features related to object-human relationships. Unlike the prior approach, the initial locations of object and human models were not known but were learned jointly from environmental features during the training process. It's essential to emphasize that the primary objective of this extended approach was to enhance 3D object detection accuracy, rather than accurately determining the locations of human models. Consequently, the inferred locations of human models may not be optimal and are heavily influenced by the presence of object types used for training the affordance models. This reliance on the training data could potentially impact the adaptability of the approach to new and previously unseen environments.

The affordance mapping process outlined in this article distinguishes itself from existing algorithms in two notable ways. First, the proposed approach frames the affordance learning process as a multi-label binary classification problem. Consequently, it can accurately assign a positive or negative label, along with a confidence value, to each grid location within the 3D scene, a capability not present in existing methods. Second, it implicitly models object-human relationships and incorporates them as features during the training phase. As a result, the model is not obligated to detect objects or their labels in order to acquire the necessary information for affordance map learning.

III. AFFORDANCE DETECTION THROUGH SUPERVISED LEARNING APPROACH

To build the affordance map, we used a supervised learning approach based on binary classification. The process of mapping unseen human action is presented in Fig.2.

Firstly, a three-dimensional point cloud map of the environment is generated. Then it is downsampled and voxalized into grid locations. For each grid location, features are extracted based on possible human skeleton models and nearby voxels. Then a multivariate structured SVM-based machine learning model is trained by utilising the retrieved features and ground truth labels. These labels are then used to map the possibility of human activity at each grid location.

Given that multiple human actions can coexist within a specific space, the corresponding affordance labels are not mutually exclusive. Therefore, the task of mapping unseen actions can be characterized as a multi-label classification problem. We partitioned each 3D image space into a four-dimensional grid, defined by the coordinates (x, y, z) of the skeleton model and the orientation θ. For each affordance type, we trained an individual binary classifier. Each binary classifier makes predictions based on a binary label vector $\bar{\mathbf{y}} = (y_1, .y_i.., y_n)$, where $y_i \in \{+1, -1\}$ for the feature vector $\bar{\mathbf{x}} = (x_1, \ldots., x_n)$, with $x_i \in \Re$. The label y_i is set to $+1$ if the grid location supports the tested affordance and -1 otherwise.

A. VIRTUAL HUMAN SKELETON MODEL FOR AFFORDANCE MAP

The proposed approach for affordance mapping models interactions between individuals and their environment using virtual humans, rather than employing real humans present in the environment. To create the necessary representations for affordance mapping, human skeleton models were acquired from a human activity detection dataset [18]. These skeleton models were obtained by recording actual people performing

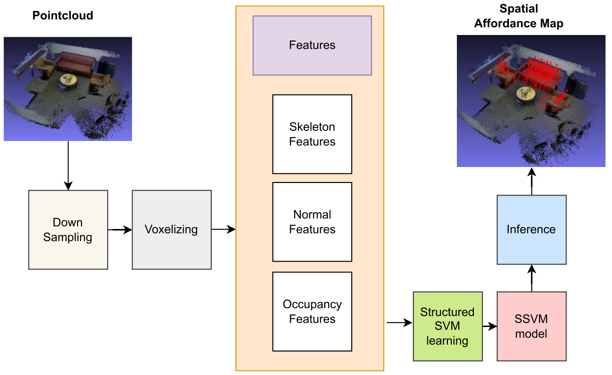


FIGURE 2. Spatial affordance mapping process.

various activities using a depth camera. The collected skeleton models were grouped into three categories through K-means clustering, with the most frequently occurring posture in each cluster illustrated in Fig.3. Each of these skeleton models corresponds to one of three affordance types: sitting-relaxing, sitting-working, and standing-working, and they are employed for affordance mapping. Each human model comprises a skeleton with 15 joint body positions located in 3D space. To create feature vectors, these virtual skeleton models are transformed to each grid location in the map defined by (x, y, z, θ). This transformation is achieved by relocating the 3D points of the human skeleton, denoted as H_l, across the given environment using rigid body transformations involving translation and rotation.

B. VOXELISATION AND SEARCH SPACE

In the initial phase of affordance mapping, the input 3D images undergo voxelization to represent grid positions. The input image is voxelized by dividing it into grids measuring 10 cm in each dimension (10 cm × 10 cm × 10 cm). The rotation θ of each skeleton model is assessed with a resolution of 0.1 radians at each grid level. To reduce the search space, the grid search along the z-axis is constrained to one grid level above and below the torso position. Despite these simplifications, the search space for a 10m x 10m room encompasses as many as 1,890,000 (100 × 100 × 3 × 63) distinct grid locations.

C. SKELETON FEATURES

This article introduces a novel set of features derived from virtual human skeleton models. These features are designed to explicitly capture the interactions between humans and their environment. The subsequent sections will detail various types of features employed in constructing the machine-learning model for the affordance map.

1) DISTANCE AND COLLISION FEATURES

The interaction between the human skeleton and the environment is modelled using the distance and collision features.

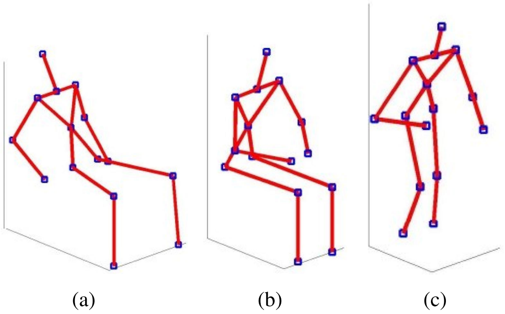


FIGURE 3. Types of affordances and their associated skeleton models. (a) Sitting-Relaxing (b) Sitting-Working (c) Standing-Working.

In order to avoid collisions between the skeleton model and environment objects, these features were chosen depending on how proximal objects were to the skeleton models.

The initial step of feature extraction involves modelling the environment through the use of a 3D Distance Transform Map, denoted as $DT(x)$, and a 3D Occupancy Map, denoted as $OC(p)$, where $\mathbf{p}$ represents any 3D position within the environment. The 3D Distance Transform (DT) serves as a shape representation that conveys the minimum distance from a point in the environment to the closest occupied voxel. In this methodology, the 3D Distance Transform is computed based on the occupied voxels within 3D point clouds, OC. The distance transform map $DT(p)$ of the occupancy grid map OC can be generated utilizing an unsigned distance function (1) that characterizes the Euclidean distance from each location $\mathbf{p}$ in the environment to the nearest occupied voxel in $OC(p)$

$$DT(p) = \min_{O_j \in OC} |O_j - p| \tag{1}$$

The human model is systematically traversed across the voxels within the surrounding area, and a distance measure is computed for each individual skeleton point of the human model. After modelling the environment as per Equation (1), we can effectively calculate distance features for a human skeleton model H at the k^{th} location and orientation $\mathbf{g}_k = (x_k, y_k, z_k, \theta_k)$ using Equation (2). Here, n represents the number of 3D points in the skeleton. The function $H(\mathbf{g}_k)$ yields a point cloud of the skeleton with n 3D points, corresponding to the transformed pose at $\mathbf{g}_k$.

$$[d_1, d_2, \ldots d_n] = DT(H(\mathbf{g}_k)) \tag{2}$$

Likewise, we can assess the possibility of collisions for a skeleton at any given location and orientation, denoted as X_k, using Equation (3). In the event of a collision, c_i is set to 1; otherwise, it is assigned a value of 0.

$$[c_1, c_2, \ldots c_n] = OC(H(\mathbf{g}_k)) \tag{3}$$

2) NORMAL FEATURES

Normal features constitute another set of features employed for affordance detection. These normal features are designed to identify the vertical and horizontal planes within the 3D point clouds. This choice of features is motivated by the fact that the majority of affordances are associated with vertical and horizontal planes.

To compute normal features, a 1m × 1m × 1m cubic volume is considered, starting from the torso position of a skeleton model at the location (x, y, z, θ). This volume is then divided into 10cm × 10cm × 10cm voxels. Subsequently, the surface normal values of points within each grid cell are averaged to derive the normal features for each voxel.

D. DATASET AND GROUND TRUTH LABELS

In order to train the affordance maps, we generated a set of dense 3D scenes using an ASUS Xtion depth camera and a 3D mapping software [19]. The resulting dataset comprises 12 high-quality 3D scenes captured in both office and domestic environments. The process of building the proposed affordance maps is essentially a supervised learning problem, necessitating the availability of ground truth labels to enable the learning of classifier parameters. Consequently, all conceivable locations within the 3D images that support the tested affordances were manually labelled for each affordance type.

IV. AN APPROACH FOR LEARNING THE AFFORDANCE MAP WITH STRUCTURED SVM ALGORITHM

Table 1 provides a summary of the number of positive and negative examples within the dataset for each affordance type. It's evident from the table that all affordance types in the dataset contain significantly more negative examples than positive ones. Effective handling of such class imbalances is crucial for any affordance detection algorithm to construct an affordance map in a large room. However, previous research has shown that classifiers often yield suboptimal results when trained on highly imbalanced datasets [20], [21]. Even common techniques like sample reweighting and negative oversampling [22] may not entirely alleviate this issue. The most effective solution is to train the classifier to optimize a metric that accurately assesses imbalanced datasets, such as the F1 score. For this purpose, Structured SVM has been successfully applied to learn to distinguish models in cases of extreme class imbalances [20], [23].

TABLE 1. Summary of the class imbalance in the dataset.

Affordance	# Positive Examples	# Negative Examples	Imabalance Ratio
Sitting-Relaxing	2457	1285326	1:523
Sitting-Working	213	2570439	1:12067
Standing-Working	391	1284935	1:3286

Unlike conventional SVMs and many other learning algorithms that primarily optimize the error rate during training, S-SVM learning has the unique capability to directly optimize performance metrics like the F1 score. This approach is particularly well-suited for imbalanced datasets. The reason this binary classification is defined as a structured problem is because the F1 score, unlike the error rate, is not solely a function of individual examples; instead, it operates as a function of a set of examples. This makes it an effective choice for addressing the challenges posed by imbalanced datasets.

V. AFFORDANCE LEARNING WITH PERFORMANCE MEASURE F1-SCORE OPTIMIZED STRUCTURED SVM

To learn affordances, we framed the learning problem as a multivariate prediction task, solvable through the S-SVM framework, with optimization based on the F1 score. Once the feature vectors $\bar{\mathbf{x}}_\mathbf{k} = (\mathbf{x_1}, \ldots, \mathbf{x_n})$ for each voxel in the 3D point cloud k and its associated label $\bar{y}_k = (y_1, \ldots, y_n)$ are known, the objective is to find a function h that maps a tuple $\bar{\mathbf{x}}_\mathbf{k} \in \bar{X}$ of n feature vectors to a tuple $\bar{y}_k \in \bar{Y}$ of n labels using Equation (4), where $\mathbf{w}$ represents the parameter vector and Ψ is the feature function that defines the relationship between features $\mathbf{x}$ and the output $\bar{y}'$.

$$h_w(\bar{\mathbf{x}}_k) = \arg\max_{\bar{y}' \in \bar{Y}} \{\mathbf{w}^T \Psi(\bar{\mathbf{x}}_k, \bar{y}')\} \tag{4}$$

The function $h_w(\bar{\mathbf{x}}_k)$ returns a set of labels $\bar{y}' = (y'_1, \ldots\ldots, y'n)$ that receive high scores according to the discriminant function. The efficient calculation of the *argmax* in equation (4) depends solely on the structure of the feature function, Ψ. When the feature function is defined as a linear combination of labels and features, as illustrated in (5), then (4) can indeed be computed efficiently.

$$\Psi(\bar{\mathbf{x}}_k, \bar{y}_k) = \sum_{i=1}^{n} y_i \mathbf{x}_i \tag{5}$$

Given that the feature function provided in (5) is linearly decomposable with respect to $\bar{y}'$, the solution can be computed by maximizing each element of $\bar{y}'$ individually. This computational approach is fast and efficient. While it is possible to explore alternative, nonlinear functions for ψ, such as those commonly used in deep learning [24], doing so would introduce a significant computational overhead. This avenue for exploration is left for future work.

One of the primary advantages of the multivariate rule $\hat{h}$ for SVM optimization is its capacity to include a loss function Δ that is based on a set of examples (e.g., F_1 score, Precision/Recall), as opposed to optimizing based on a single example-based loss function like the error rate. This enables the algorithm to effectively consider the performance across multiple examples when making predictions.

A. OPTIMIZATION PROBLEM

To train the discriminant $\bar{h}_w$ function, we used the following generalization of the support vector machine.

$$\begin{aligned} &\min_{w,\xi} \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m} \xi_k \\ &\text{s.t. } \xi_k \geq 0, \quad \forall k \\ &\quad \mathbf{w}^T[\Psi(\bar{\mathbf{x}}_k, \bar{y}_k) - \Psi(\bar{\mathbf{x}}_k, \bar{y}'] \geq \Delta(\bar{y}', \bar{y}_k) - \xi_k, \quad \forall k, \\ &\quad \forall \bar{y}' \in \mathcal{Y} \setminus \bar{y}_k \end{aligned} \tag{6}$$

The loss function $\bigtriangleup(\bar{y}', \bar{y}k)$ is a measure that decreases during training as the predicted tuple of outputs $\bar{y}'$ approaches the ground truth label tuple $\bar{y}k$. This optimization is convex; however, unlike the standard SVM, the above optimization problem involves an infeasibly large number of constraints, equivalent to the number of training examples multiplied by each possible tuple of $\bar{y}_k \in \mathcal{Y}$. This results in an intractable optimization problem. Nevertheless, the use of the sparse approximation algorithm proposed by [25] allows for the solution of this optimization problem in polynomial time for many types of multivariate loss functions.

The objective function that needs to be minimized in (6) represents a trade-off between the model's complexity, denoted as $\| \mathbf{w} \|$, and the hinge loss relaxation of the training loss, $\sum \xi_k$. Here, $C > 0$ is a constant that controls this trade-off.

B. MAXIMIZATION STEP AT INFERENCE

The efficiency of the proposed S-SVM method is significantly contingent on the effective computation of *argmax* operations within the inference equation (4). During the inference step, we need to calculate:

$$\bar{h}_w(\bar{\mathbf{x}}_k) = \arg\max_{\bar{y}' \in \bar{Y}} \{\mathbf{w}^T \Psi(\bar{\mathbf{x}}_k, \bar{y}')\} \tag{7}$$

By substituting the feature function (5) into (7), we can obtain the following form of the inference equation.

$$\bar{h}_w(\bar{\mathbf{x}}_k) = \arg\max_{\bar{y}' \in \bar{Y}} \{\mathbf{w}^T \sum_{i=1}^{n} y'_i \mathbf{x}_i\} \tag{8}$$

As the feature function is linearly decomposable with respect to a single label, y'_i, the above inference equation can be simplified into the following form, as shown below.

$$\bar{h}_w(\bar{\mathbf{x}}_k) = \arg\max_{\bar{y}' \in \bar{Y}} \{\sum_{i=1}^{n} \mathbf{w}^T y'_i \mathbf{x}_i\} \tag{9}$$

The *argmax* operation in (9) can be further decomposed into a single feature vector and a single label, as illustrated below. This decomposition renders the structured SVM prediction problem equivalent to the conventional SVM prediction problem

$$*y'_i = \arg\max_{y'_i \in \bar{y}'} \{\mathbf{w}^T y'_i \mathbf{x}_i\}, \quad \forall y'_i \in \bar{y}' \tag{10}$$

As each label is a binary value with $y'_i = \{+1, -1\}$ computing prediction becomes very efficient with

$$\begin{aligned} \bar{h}_w(\bar{\mathbf{x}}_k) &= \arg\max_{\bar{y}' \in \bar{Y}} \{\mathbf{w}^T \Psi(\bar{\mathbf{x}}_k, \bar{y}')\} \\ &= (sign(\mathbf{w}.\mathbf{x}_1), \ldots\ldots.., sign(\mathbf{w}.\mathbf{x}_n)) \end{aligned} \tag{11}$$

Hence, the linear feature function Ψ facilitates rapid predictions by significantly reducing the size of the output space $(\bar{Y})$ from 2^n to n, where n represents the number of distinct grid locations in the map.

C. MAXIMIZATION STEP AT LEARNING

Calculating the loss-augmented *argmax* required for training is a more complex process, and its nature depends on the specific loss function in use. Therefore, the maximization step during training requires a sophisticated algorithm to enable efficient training.

The step that computes the most violated constraint in the cutting plane optimization algorithm must complete the following maximization step, as shown in equation (12).

$$\bar{y}' \leftarrow \arg\max_{\bar{y}' \in \bar{Y}} \{\bigtriangleup(\bar{y}', \bar{y}_k) + \mathbf{w}^T \Psi(\bar{\mathbf{x}}_k, \bar{y}')\} \tag{12}$$

If the loss function $\bigtriangleup(\bar{y}', \bar{y}_k)$ is linear in $\hat{y}'$ (e.g., Hamming Loss), then the solution for (12) can be computed by maximizing $\bar{y}'$ element-wise. However, linear loss functions like the Hamming loss may struggle with the label bias issue associated with highly imbalanced classes. To address this problem, it is advisable to directly optimize for performance measures such as the F_1 score, which is the harmonic mean of precision and recall. This can be achieved within the structural SVM framework by incorporating a loss function based on the contingency table in the maximization step.

1) LOSS FUNCTION BASED ON CONTINGENCY TABLE

When the loss function becomes nonlinear in $\bar{y}'$, computing the *argmax* in (12) becomes challenging, as an exhaustive search over all possible $\bar{y}'$ is infeasible. However, the computation of the *argmax* can be categorized over all possible contingency tables, allowing each sub-problem to be solved efficiently.

The contingency table for binary classification is illustrated in Table 2, where a represents the number of true positives, b represents the number of false positives, c represents the number of false negatives, and d represents the number of true negatives.

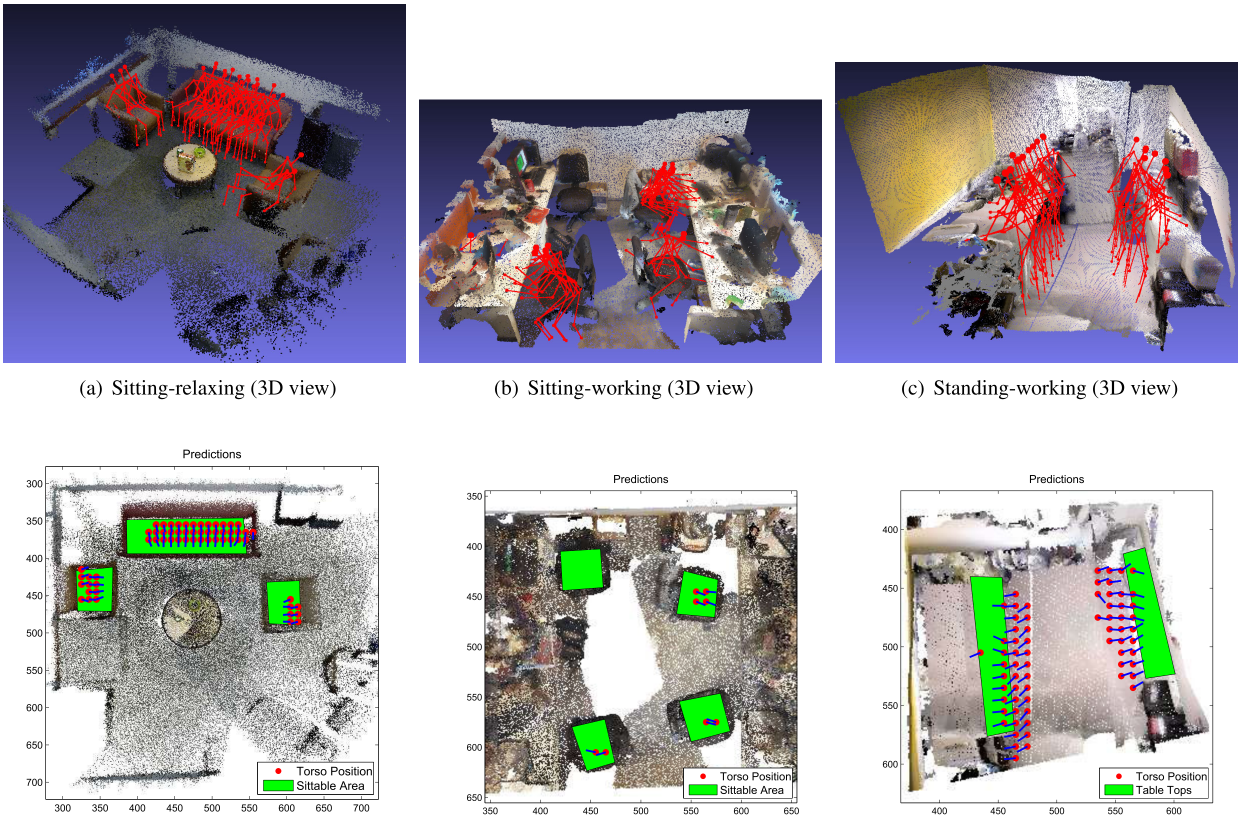
TABLE 2. Contingency table for binary classification.

	y=1	y=-1
h(x)=1	a	b
h(x)=-1	c	d

It is evident that there are only on the order of $O(n^2)$ different contingency tables for a binary classification problem with n samples. Therefore, if the loss function $\bigtriangleup(a, b, c, d)$ is computed based on Table 2, it can have at most $O(n^2)$ distinct values. With the loss function $\bigtriangleup(a, b, c, d)$, the *argmax* for calculating the most violated constraint in (6) can be redefined as shown in (13).

$$\bar{y}' \leftarrow \arg\max_{\bar{y}' \in \bar{\mathcal{Y}}} \{\bigtriangleup(a, b, c, d) + \mathbf{w}^T \Psi(\bar{\mathbf{x}}_k, \bar{y}')\} \tag{13}$$

The computation of *argmax* in (13) is exhaustive, but its efficiency can be improved by organizing the search space of $\bar{\mathcal{Y}}$ across different contingency tables. In [25], a novel algorithm for computing *argmax* with loss functions derived from the contingency table is introduced. It calculates *argmax*



(a) Sitting-relaxing (3D view) (b) Sitting-working (3D view) (c) Standing-working (3D view)

(d) Sitting-relaxing (Planview) (e) Sitting-working (Planview) (f) Standing-working (Planview)

FIGURE 4. Qualitative results of the S-SVM learning for affordance mapping. The top row presents the inferred 3D skeletons by the S-SVM model, while the bottom row showcases the top-down view of the skeletons along with their orientations.

over all $\overline{\mathcal{Y}}_{abcd}$ for each contingency table (a, b, c, d), which is a subset of $\bar{y}'$ for the selected contingency table, as shown in (14) and (15).

$$\bar{y}_{abcd} = \arg \max_{\bar{y}' \in \overline{\mathcal{Y}}_{abcd}} \{\mathbf{w}^T \Psi(\overline{\mathbf{x}_k}, \bar{y}')\} \tag{14}$$

$$\bar{y}_{abcd} = \arg \max_{\bar{y}' \in \overline{\mathcal{Y}}_{abcd}} \{\mathbf{w}^T \sum_{i=1}^{n} y_i' \mathbf{x}_i\} \tag{15}$$

Since the inference function in (15) is linear with respect to $\bar{y}'$, the solution can be obtained by maximizing $\bar{y}'$ element-wise. The maximum value of the objective function for a particular contingency table is achieved when the a positive examples with the largest values of $(\mathbf{w}^T \mathbf{x}_i)$ are classified as positive, and the d negative examples with the lowest values of $(\mathbf{w}^T \mathbf{x}_i)$ are classified as negative. Ultimately, the overall *argmax* can be determined by maximizing over each of these maxima along with their corresponding loss function values.

VI. RESULTS

This section presents the experimental results of the affordance mapping process based on the S-SVM algorithm. Since structured output SVM is a supervised learning algorithm, it necessitates training before inferring the affordance map in a new environment. After computing the feature set for each image in the training dataset, the cutting plane algorithm described in [25] is employed to calculate the weight vector, $\mathbf{w}$. Subsequently, the inference function (11) is used to predict the affordance map in a new 3D image.

It's important to note that the trained parameters can potentially overfit the training data because each iteration of the proposed structured output SVM algorithm aims to improve the F1 score. This could lead to a decrease in performance on an unknown input image. To mitigate the overfitting problem, a validation set is employed. The best parameter set for S-SVM training is determined by recording the parameter set for each iteration of the training process and then testing it on a validation set. For inference, the parameter set that yields the best classification results on the validation set is utilized. The classifier's performance is evaluated using the k-fold cross-validation method.

A. QUALITATIVE ANALYSIS

We conducted a qualitative analysis of affordance mapping by visualizing the skeleton model of detected affordances on the grid locations of the 3D point clouds from the test dataset. Fig. 4, shows the qualitative results for each affordance type that utilises sample point clouds from the test set. The top row displays the 3D skeleton view, while the bottom row presents the plan view. The red circles denote the torso, and the associated blue lines indicate the orientation of each skeleton in the plan view. These qualitative results clearly demonstrate the successful application of S-SVM learning in identifying the grid locations that support each affordance type. Additionally, the results show the correct orientation of each skeleton associated with each affordance type.

It's worth noting that a few outliers were observed when testing for the sitting-working and standing-working affordance types. This is expected, as there are extreme class imbalances in those affordance types compared to the sitting-relaxing affordance type. Even with such extreme class imbalances in the dataset, the S-SVM-based learner was able to produce acceptable results.

TABLE 3. Comparison of performance measures in the k-fold cross-validation test.

Affordance Type	SVM Algorithm	Precision	Recall	F1
Sitting Relaxing	Random Sampling	0.36	0.33	0.35
	Focused Sampling	0.32	0.37	0.35
	Z-SVM	0.51	0.35	0.35
	Different Cost Model	0.28	0.44	0.34
	S-SVM	0.65	0.80	**0.72**
Standing Working	Random Sampling	0.09	0.49	0.13
	Focused Sampling	0.17	0.30	0.22
	Z-SVM	0.12	0.31	0.15
	Different Cost Model	0.10	0.42	0.14
	S-SVM	0.28	0.81	**0.42**
Sitting Working	Random Sampling	0.55	0.52	0.50
	Focused Sampling	0.34	0.51	0.38
	Z-SVM	0.27	0.84	0.37
	Different Cost Model	0.50	0.55	**0.52**
	S-SVM	0.40	0.77	0.48

B. QUANTITATIVE ANALYSIS

The comparison of average performance measures of k-fold cross-validation between S-SVM and other SVM algorithms modified to handle class imbalances is shown in Table 3. We compared the results of four existing SVM learners, namely: Random Sampling [26], Focused re-sampling [27], Z-SVM [28], and Different Cost model SVM [29] with the S-SVM method proposed in this article. The average precision and recall values of the best F1-score recorded in the class imbalance test for each affordance type are shown in Table 3.

The S-SVM method achieved significantly higher F1 scores in the affordance types Sitting-Relaxing and Standing-Working than the next best F1 score in each category. It has the second-best F1 score in Sitting-Working affordance, with a difference of only -0.04 from the highest recorded F1-score in that category. Overall, the S-SVM method performed consistently in all three affordance types. This is due to the fact that, unlike other methods, S-SVM was trained directly optimising on the F1 score. As a result, the class imbalance problem has little impact on it.

VII. CONCLUSION AND PROSPECTS

This article presented the framework of learning spatial affordances using 3D point clouds for mapping unseen human actions in indoor environments. The key outcomes of this article are summarized as follows:

- The problem of learning human context in indoor environments was addressed by proposing an affordance map that analyses geometric characteristics of the environment to predict potential affordances in the environment as human skeleton models with affordance likelihoods.
- The S-SVM framework was optimised on the performance metric F1-score for addressing the multivariate prediction problem that was devised to learn affordances. This method defined the process of creating an affordance map as a structured problem that could be learned efficiently even with extremely imbalanced data.
- Qualitative analysis observationÂ clearly demonstrated that S-SVM learning is effectively used to determine the grid location that supports each affordance type. Further, the evaluation showedÂ each skeleton's correct orientationÂ in relation to each affordance type. The S-SVM-based learner delivered satisfactory results even with the dataset's high-class imbalances.
- On average the proposed S-SVM classifier outperformed other previously reported classifiers based on quantitative evaluation results. The proposed S-SVM method outperformed all other models for the sitting-relaxing and standing-working affordance types, but it only obtained a -0.04 F1-score over the best model for the sitting-working affordance type.

The spatial affordance map holds the potential for enhancing various aspects of future work, including improving existing human activity recognition algorithms, social-aware path planning algorithms, and active object search algorithms, as described in previous research [30], [30], [31].

Most current activity recognition algorithms rely on having a full-body view of humans to accurately recognize activities. However, in cluttered environments, achieving this full view can be challenging due to occlusions. In such scenarios, the affordance map can be leveraged to predict locations that offer unobstructed views of humans. This, in turn, can assist robots in more accurately tracking human body parts during activity recognition.

Furthermore, the data from the affordance map can also enhance human detection and tracking algorithms. The information on the human skeleton provided by the affordance map can be particularly valuable for active object search in indoor environments. These applications highlight the potential significance of affordance maps in advancing human-robot interaction and autonomous robotics.

The affordance process proposed in this paper, based on S-SVM, is constrained by the manual engineering of features. However, the incorporation of deep learning algorithms with automatic feature selection has the potential to enhance the accuracy of affordance mapping.

REFERENCES

[1] D. Wei, L. Chen, L. Zhao, H. Zhou, and B. Huang, "A vision-based measure of environmental effects on inferring human intention during human robot interaction," *IEEE Sensors J.*, vol. 22, no. 5, pp. 4246–4256, Mar. 2022, doi: 10.1109/JSEN.2021.3139593.

[2] S. Sehestedt, S. Kodagoda, and G. Dissanayake, "Robot path planning in a social context," in *Proc. IEEE Conf. Robot., Autom. Mechatronics*, Jun. 2010, pp. 206–211.

[3] L. Piyathilaka and S. Kodagoda, "Human activity recognition for domestic robots," in *Field and Service Robotics: Results of the 9th International Conference on Field and Service Robotics FSR held at Brisbane, Australia*, L. Mejias, P. Corke, and J. Roberts, Eds. Cham, Switzerland: Springer, 2015, pp. 395–408.

[4] L. Piyathilaka and S. Kodagoda, "Gaussian mixture based HMM for human daily activity recognition using 3D skeleton features," in *Proc. IEEE 8th Conf. Ind. Electron. Appl. (ICIEA)*, Jun. 2013, pp. 567–572.

[5] Y. Jiang, H. Koppula, and A. Saxena, "Hallucinated humans as the hidden context for labeling 3D scenes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 2993–3000.

[6] E. J. Gibson and A. D. Pick, *An Ecological Approach to Perceptual Learning and Development*. London, U.K.: Oxford Univ. Press, 2000.

[7] L. Piyathilaka and S. Kodagoda, "Affordance-map: Mapping human context in 3D scenes using cost-sensitive SVM and virtual human models," in *Proc. IEEE Int. Conf. Robot. Biomimetics (ROBIO)*, Dec. 2015, pp. 2035–2040.

[8] S. C. Mukhopadhyay, "Wearable sensors for human activity monitoring: A review," *IEEE Sensors J.*, vol. 15, no. 3, pp. 1321–1330, Mar. 2015, doi: 10.1109/JSEN.2014.2370945.

[9] Y. Zhang, "Detection and tracking of human motion targets in video images based on camshift algorithms," *IEEE Sensors J.*, vol. 20, no. 20, pp. 11887–11893, Oct. 2020, doi: 10.1109/JSEN.2019.2956051.

[10] Q. Hao, D. J. Brady, B. D. Guenther, J. B. Burchett, M. Shankar, and S. Feller, "Human tracking with wireless distributed pyroelectric sensors," *IEEE Sensors J.*, vol. 6, no. 6, pp. 1683–1696, Dec. 2006, doi: 10.1109/JSEN.2006.884562.

[11] C. Will, P. Vaishnav, A. Chakraborty, and A. Santra, "Human target detection, tracking, and classification using 24-GHz FMCW radar," *IEEE Sensors J.*, vol. 19, no. 17, pp. 7283–7299, Sep. 2019, doi: 10.1109/JSEN.2019.2914365.

[12] H. Li, A. Shrestha, H. Heidari, J. Le Kernec, and F. Fioranelli, "Bi-LSTM network for multimodal continuous human activity recognition and fall detection," *IEEE Sensors J.*, vol. 20, no. 3, pp. 1191–1201, Feb. 2020, doi: 10.1109/JSEN.2019.2946095.

[13] F. Luo, S. Poslad, and E. Bodanese, "Human activity detection and coarse localization outdoors using micro-Doppler signatures," *IEEE Sensors J.*, vol. 19, no. 18, pp. 8079–8094, Sep. 2019, doi: 10.1109/JSEN.2019.2917375.

[14] L. Lu, H. Wang, B. Reily, and H. Zhang, "Robust real-time group activity recognition of robot teams," *IEEE Robot. Autom. Lett.*, vol. 6, no. 2, pp. 2052–2059, Apr. 2021, doi: 10.1109/LRA.2021.3060723.

[15] L. Fiorini, F. G. C. Loizzo, A. Sorrentino, J. Kim, E. Rovini, A. Di Nuovo, and F. Cavallo, "Daily gesture recognition during human-robot interaction combining vision and wearable systems," *IEEE Sensors J.*, vol. 21, no. 20, pp. 23568–23577, Oct. 2021, doi: 10.1109/JSEN.2021.3108011.

[16] H. Grabner, J. Gall, and L. V. Gool, "What makes a chair a chair?" in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 1529–1536.

[17] Y. Jiang, M. Lim, and A. Saxena, "Learning object arrangements in 3D scenes using human context," 2012, *arXiv:1206.6462*.

[18] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Unstructured human activity detection from RGBD images," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2012, pp. 842–849.

[19] I. Dryanovski, R. G. Valenti, and J. Xiao, "Fast visual odometry and mapping from RGB-D data," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2013, pp. 2305–2310.

[20] Y. Tang, Y.-Q. Zhang, N. V. Chawla, and S. Krasser, "SVMs modeling for highly imbalanced classification," *IEEE Trans. Syst. Man, Cybern. B, Cybern.*, vol. 39, no. 1, pp. 281–288, Feb. 2009.

[21] R. Akbani, S. Kwek, and N. Japkowicz, "Applying support vector machines to imbalanced datasets," in *Machine Learning: ECML 2004*. Pisa, Italy: Springer, 2004, pp. 39–50.

[22] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, pp. 321–357, Jun. 2002.

[23] Y. Li, D. J. Crandall, and D. P. Huttenlocher, "Landmark classification in large-scale image collections," in *Proc. IEEE 12th Int. Conf. Comput. Vis.*, Sep. 2009, pp. 1957–1964.

[24] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[25] T. Joachims, "A support vector method for multivariate performance measures," in *Proc. 22nd Int. Conf. Mach. Learn. (ICML)*, 2005, pp. 377–384.

[26] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[27] R. Batuwita and V. Palade, "Efficient resampling methods for training support vector machines with imbalanced datasets," in *Proc. Int. Joint Conf. Neural Netw. (IJCNN)*, Jul. 2010, pp. 1–8.

[28] T. Imam, K. M. Ting, and J. Kamruzzaman, "z-SVM: An SVM for improved classification of imbalanced data," in *AI 2006: Advances in Artificial Intelligence*. Hobart, TAS, Australia: Springer, 2006, pp. 264–273.

[29] T. Joachims, "Transductive inference for text classification using support vector machines," in *Proc. ICML*, vol. 99, 1999, pp. 200–209.

[30] L. Piyathilaka and S. Kodagoda, "Affordance-map: A map for context-aware path planning," in *Proc. Australas. Conf. Robot. Automat. (ACRA)*, 2014.

[31] L. Piyathilaka, "Learning hidden human context in 3D office scenes by mapping affordances through virtual humans," *Unmanned Syst.*, vol. 3, no. 4, pp. 299–310, 2015.

LASITHA PIYATHILAKA (Member, IEEE) received the B.Sc. (Eng.) degree (Hons.) in electrical engineering, in 2004, the M.Phil. degree from the University of Moratuwa, Sri Lanka, in 2011, and the Ph.D. degree in robotics from the University of Technology Sydney (UTS), in 2016.

With years of industry experience, he has actively contributed to various university-industry collaboration projects. Currently, he is a Lecturer with Central Queensland University (CQU). Previously, he held the position of a Research Fellow with UTS. During his tenure with UTS, he led an engineering team dedicated to developing sensing and robotic systems for inspecting underground infrastructure. His expertise in robotics and engineering has been instrumental in driving advancements in this field.

SARATH KODAGODA (Member, IEEE) received the B.Sc. (Eng.) degree (Hons.) in electrical engineering from the University of Moratuwa, Sri Lanka, in 1995, and the M.Eng. and Ph.D. degrees in robotics from Nanyang Technological University, Singapore, in 2000 and 2004, respectively. He is currently a Professor and the Director of the UTS Robotics Institute, the Founder of the iPipes Laboratory, and a Program Coordinator of the Mechanical and Mechatronics Degree with the University of Technology Sydney, Australia. He is also the President of the Australian Robotics and Automation Association. His current research interests include infrastructure robotics, sensors and perception, machine learning, and human–robot interaction. His works in in-pipe sensing and robotics has won several awards, including the 2020 National Research Innovation Award from the Australian Water Association.

KARTHICK THIYAGARAJAN (Senior Member, IEEE) received the B.E. degree in electronics and instrumentation engineering from Anna University, Chennai, India, in 2011, the M.Sc. degree in mechatronics from Newcastle University, Newcastle upon Tyne, U.K., in 2013, and the Ph.D. degree in smart sensing from the University of Technology Sydney (UTS), Sydney, Australia, in 2018.

He is currently a Research Fellow with the UTS Robotics Institute, UTS. His current research interests include the development of intelligent sensing and perception technologies for robotics.

Dr. Thiyagarajan is a Secretary of the IEEE Sensors Council NSW Chapter, in 2023 and 2022, and an Executive Committee Member of the IEEE Sensors Council Young Professionals and Publicity Committee, in 2022.

MASSIMO PICCARDI (Senior Member, IEEE) received the M.Eng. and Ph.D. degrees from the University of Bologna, Bologna, Italy, in 1991 and 1995, respectively. He is currently a Full Professor in computer systems with the University of Technology Sydney, Australia. His research interests include natural language processing, computer vision, and pattern recognition. He has coauthored over 150 articles in these areas. He is a member of the IEEE Computer and Systems, Man, and Cybernetics Societies, and a member of the International Association for Pattern Recognition. He serves as an Associate Editor for IEEE Transactions on Big Data.

D. M. G. PREETHICHANDRA (Senior Member, IEEE) received the B.Sc. (Eng.) degree in electrical and electronics engineering from the University of Peradeniya, Sri Lanka, the M.Eng. degree in telecommunications and the Ph.D. degree in sensor design from Saga University, Japan, and the master's degree in environmental and business management from The University of Newcastle, Australia. He was an Assistant Professor in Japan. He is currently an Associate Professor and the Head of the Discipline of Mechatronics Engineering and Central Queensland University, Australia. He has been working in academia for more than 30 years and working on nano-biosensor development for environmental and biomedical measurements, sensor networking, robotics, and embedded systems design.

UMER IZHAR (Member, IEEE) received the Ph.D. degree in electrical engineering, with a focus on micro-electro-mechanical-systems (MEMS) from Lehigh University, PA, USA, through the Fulbright Scholarship. After completing the Ph.D. degree, he was an Assistant Professor with NUST, Pakistan. Later, he joined Central Queensland University (CQU), Australia, as a Lecturer in mechatronics engineering and developed part of the mechatronics curriculum. He is currently a Lecturer and a Program Coordinator in mechatronics engineering with UniSC. He has held many academic leadership and coordination positions throughout his career during which he contributed to the development of education. He has a growing research profile which includes conference and journal publications. His current research interests include the design and fabrication of MEMS-based sensors and actuators, robotics, and automation.