Gesture Recognition Intermediary Robot for Abnormality Detection in Human Activities

Salisu Wada Yahaya*, Ahmad Lotfi*, Mufti Mahmud*, Pedro Machado* and Naoyuki Kubota†
*Nottingham Trent University
School of Science and Technology, Clifton Lane, NG11 8NS, United Kingdom
Email: salisu.yahaya2015@my.ntu.ac.uk; {ahmad.lotfi, mufti.mahmud, pedro.baptistamachado}@ntu.ac.uk
†Tokyo Metropolitan University
Graduate School of System Design, 6-6 Asahigaoka, Hino, Tokyo 191-0065, Japan
Email: kubota@tmu.ac.jp

***Abstract*—The world ageing population is increasing, giving rise to research targeted towards improving the quality of life and promoting the independent living of older adults. Detecting abnormalities in the daily activities of the older adults is relevant since abnormalities can be an early sign of health decline, prompting for the need for intervention. Current approaches to abnormality detection involve modelling the usual behavioural routine of the individual as a baseline and comparing subsequent behaviour to the baseline to detect abnormalities. This approach is prone to errors and not flexible since it does not take into account changes in human behavioural routine. Training is usually performed on pre-existing data making the abnormality detection model non-adaptive to new incoming data. An intermediary can be incorporated to enable model predictions to be communicated to humans for verification of the detected anomalies. This paper proposes a gesture recognition approach for facilitating interaction between humans and a robot intermediary. A model capable of recognising hand gestures corresponding to affirmations and denials is implemented. Preliminary evaluation shows that the proposed gesture recognition approach has the potential of being utilised in an assistive robot intermediary.**

***Keywords*—Gesture Recognition, Assistive Robot, Anomaly Detection, Abnormality Detection, Activities of Daily Living (ADL), Convolutional Neural Network (CNN), You Only Look Once (YOLO)**

I. Introduction

Assistive robots have the potential of being utilised in a home environment for various purposed ranging from providing domestic services, companionship and monitoring. This paper explores the means of utilising an assistive robot platform to serve as an intermediary between humans and an abnormality detection system for Activities of Daily Living (ADL) of older adults. ADL are activities an individual must be able to perform independently without requiring assistance such as eating, mobility, maintaining continence and personal hygiene etc. This is driven by the need to improve the quality of life and promote independent living of the increasing ageing population which is estimated to be over 1.9 billion by 2050 [1].

Abnormality, also known as an anomaly in this context is defined as any significant deviation from the usual behavioural routine of an individual. Abnormalities in ADL can be detrimental to well-being and in most cases attributed to health decline. For example, early symptoms of Mild Cognitive Impairment (MCI) such as Dementia in older adults can be identified from changes in their routine such as frequently interrupted sleep, performing less activity during the day and much activities at night time, confusion or forgetfulness etc [2]. To detect these abnormalities, the usual behavioural routine of the individual is learned to serve as a baseline. Subsequent behaviours are then compared to the baseline to detect deviation which could be considered as abnormal. Different computational approaches for learning the usual behavioural routine of an individual and detecting abnormalities in it has been proposed and applied for the detection of various anomalies [2]–[5]. The drawback of these approaches is that they do not take into account changes in the routine of an individual. Human behaviour is dynamic and subject to changes due to factors such as social, health and seasonal influences. Consequently, the existing approaches are not able to adapt to changes in the behavioural routine and therefore, lead to the generation of high false alarm rate undermining the effectiveness of the system and lack of acceptability by the users and carers [6].

In [7], we proposed an approach for addressing this shortcoming that involves incorporating an intermediary into the anomaly detection system as shown in Figure 1. The proposed framework allows activities detected as abnormal by the computational model to be communicated to humans through the robot intermediary. The human response collected by the intermediary which can be an affirmation or denial of the model's prediction is fed back into the computational model to adapt to the user's feedback and learn incrementally. The choice of an assistive robot as an intermediary over screen-based interfaces is due to robot's support for multi-modal interaction such as through voice, touch, gesture etc and the presence of physical embodiment which facilitates interaction according to several studies [8], [9].

This paper focuses on human gesture recognition from 2D images to be utilised on the robot intermediary. The interpreted hand gestures corresponding to affirmations or denials will serve as an input to the anomaly detection model confirming the model's prediction. Gestures corresponding to an affirmation signify that the predicted activity by the computational model is not an anomaly, prompting the model to

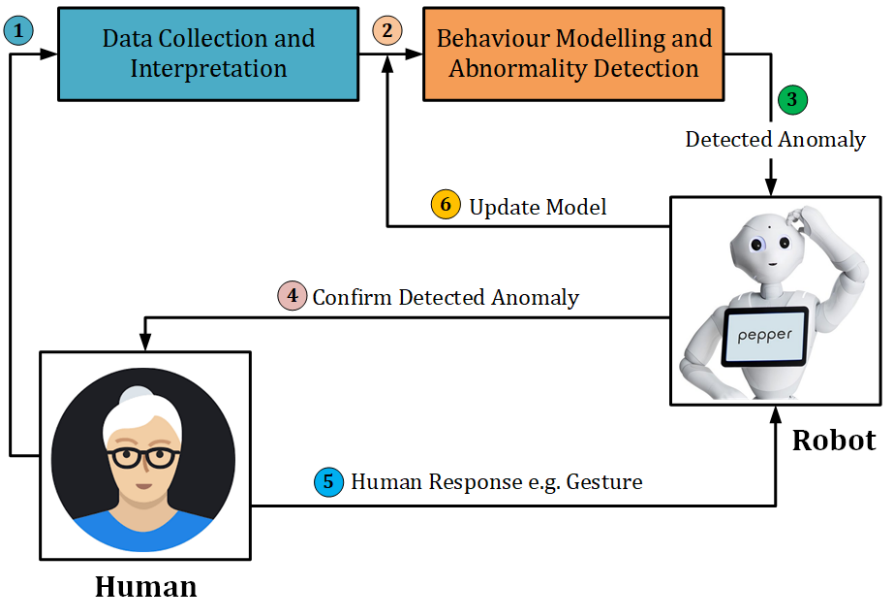

Fig. 1. A schematic diagram of anomaly detection system with a robot intermediary.

learn the characteristic features of the activity and vice-versa. The translated gesture in combination with other interaction modalities such as voice and touch screen input can be fused to achieve better Human-Robot Interaction (HRI) in this context.

The rest of the paper is organised as follows: Section II reviews related work in the area of assistive robots and gesture recognition. Section III contains the methodology while Section IV consists of the experimental results and findings. Section V provides a summary of the concluded work and future research direction.

II. Related Work

Over the years, research aimed at utilising robots for various assistive purposes has been proposed ranging from companionship, health care monitoring, motivational coach and other domestic services. A companion robot capable of helping older adults suffering from MCI to keep track of their reminders (such as medication times), establishing communication with carers and family, as well as administering exercises to improve their cognitive abilities is proposed [10]. A robot is built in [11] to administer and coach older adults in performing physical activities. It utilises Microsoft Kinect sensor to analyse human pose to compare with the predefined exercises and provide feedback in the form of speech and facial expression. Assistive robots have also been utilised to promote engagement and to serve as a means of teaching. For example, in [12], a robot platform is used to serve as a teaching tool to help diabetic children better understand how to manage their condition, while in [13], a similar platform is used to facilitate engagement in children with Autism Spectrum Condition (ASC) and to enable them to see things from other peoples perspectives also known as Visual Perspective Taking (VPT).

Research has been conducted with the aim of incorporating robots with gesture recognition capabilities for various purposes. While some research requires the use of specialised hardware such as sensor-equipped gloves or wrist-worn devices equipped with an accelerometer and gyroscope [14], the focus will be on approaches that utilise only images and video stream data. Luo et al. [15] proposed a robot for recognising hand gestures correspond to sign languages using a combination of 2D camera and a Kinect depth sensor. One of the aims among other things is that, by understanding sign language, the robot can assist disabled (deaf) people in a hospital environment with enquiries, calling a doctor, navigating the environment, and serve as an intermediary for communicating with people that do not understand sign language. To provide better HRI, the author in [16] implemented a gesture recognition system for a mobile robot. Six hand gestures are detected using images captured from the robot's 2D camera. Similarly, in [17], a dynamic pointing gesture recognition approach is proposed for utilisation in HRI. The idea is to promote a naturalistic interaction by allowing humans to control the robot's movement by pointing at a direction. The mobile robot tracks the hand movement using a stereo camera as the user get in its line of sight. Unlike in static gestures where the recognition is performed on a single image frame, dynamic gestures require the processing of multiple image frames over time to be able to track the movements thereby requiring the use of temporal models. The author applied Hidden Markov Model (HMM) for the hand movement tracking to ascertain the direction the user is pointing at.

Different computational methodologies for the recognition of various actions and gestures from camera-based data exist. Deep learning models such as Convolutional Neural Network (CNN) is the most favoured due to its ability to learn features

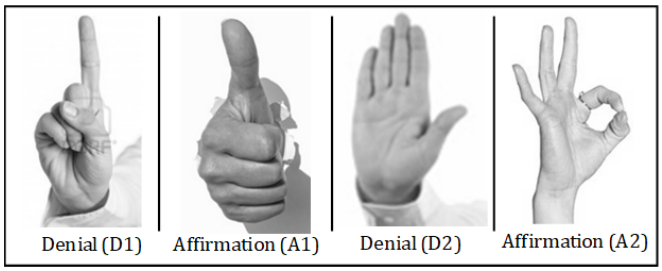

Fig. 2. Hand gestures corresponding to affirmations and denials.

encoding directly from data without requiring explicit feature engineering. It has been applied for static hand gesture recognition in [16], [18]. For dynamic gesture recognition, different deep learning approach has been proposed. One approach involves using the combination of CNN and Recurrent Neural Network (RNN) for the prediction. The CNN learns the encoding of the various images over the ordered sequences of frames while RNN learns the sequential pattern of the ordered frames [19]. Another approach involves using a 3D CNN. A 3D CNN model allows features to be extracted in both spatial and temporal dimensions by performing 3D convolution. This enables motion information encoded in various frames to be captured [20]. Traditional classification approaches have also been utilised. However, unlike in CNN, complex pre-processing of the data is required such as image cropping and resizing, background removal, feature extraction and normalisation etc. The output of the preprocessing step is supplied to a classification algorithm for the recognition of the action/gesture of interest. This approach can be combined with temporal models to allow for the recognition of dynamic gestures as applied in [16].

III. Methodology

To enable communication through the robot intermediary, 4 hand gestures are defined with 2 of the gestures corresponding to an affirmation and the remaining gestures corresponding to denial as shown in Figure 2. Detection of one of the defined gestures during the interaction indicates the human response to the query administered by the robot intermediary. Because the gestures are static in nature, the most feasible approach is to detect them from image frames. Due to the difficulty in determining the exact interval between the user's feedback (e.g. in form of gesture) and the time the query is administered by the intermediary, it is nearly impossible to identify the exact image frame containing the gestures.

The schematic diagram of the proposed gesture recognition approach is shown in Figure 3. The procedure involves first recording a short video stream (e.g. 5-15 seconds long) of the human from the period the query is administered by the robot intermediary. The recorded stream is then converted into frames of images. A computational model trained to detect the gestures of interest can then be applied to the extracted images. Uniquely identified gestures from the image frames are aggregated. Frames containing no gesture are ignored and considered as frames of images before or after the gestures are performed. The aggregated gestures detected from the image frames are used to predict the human's response with the gesture having the highest number occurrences taken as the final prediction.

CNN based model for object detection known as YOLOv3 is utilised for detection of the gestures depicted in Figure 2. The choice of YOLOv3 over other object detection methods such as Single Shot Detector (SDD), Region-based CNN (R-CNN), Fast R-CNN, Faster R-CNN etc. is due to YOLOv3 superior processing speed. Fast processing speed in the gesture recognition component is important since it allows for the realisation of a near real-time anomaly detection system. Despite having less prediction accuracy compared to the aforementioned models, the performance difference between YOLOv3 and the other models is quite negligible in this context.

A. YOLO

You Only Look Once (YOLO) is an implementation approach of CNN for detection of objects in images proposed by Redmon et al. [21]. Over the years, different versions of YOLO has been developed and recently, the third version known as YOLOv3 is released with several performance improvements compared to its predecessors. In YOLO, a single CNN trained on full images is used to make predictions of the multiple objects specified simultaneously. The input image is divided into an M x M grid of equal sizes. Each of the grid cells predicts B bounding boxes around the detected objects of interest along with their respective confidence scores. The bounding boxes each consist of 5 values corresponding to the x and y coordinates representing the centre of the bounding box, the width and height of the bounding box relative to the image dimension, and the confidence score respectively.

YOLO architecture depicted in Figure 4 consists of 24 convolutional layers for extracting features from the image and 2 fully connected layer for making a prediction and generating the bounding boxes. More information on YOLO and its implementation details can be found in [21]

B. Data Collection Scenario

Data is collected from willing participants for training and testing. Five (5) middle-aged individuals of different skin colour stood in from of a 2D camera in a controlled lab environment to perform the selected gestures. We choose a 2D gesture recognition approach because, in a real-life scenario, the robot intermediary may not necessarily be equipped with a depth camera. Moreover, using a 2D camera gives room for the utilisation of other non-robotic intermediaries (such as screen-based interfaces) as long as they are equipped with a camera. In the first scenario referred to as "Scenario 1", each of the selected gesture in Figure 2 is performed for approximately 1 minute by each participant, resulting in a 4-minute recorded video for the 4 different gestures combined per participant. The participants are asked to perform the gestures in different variations by slightly rotating their hands, changing their

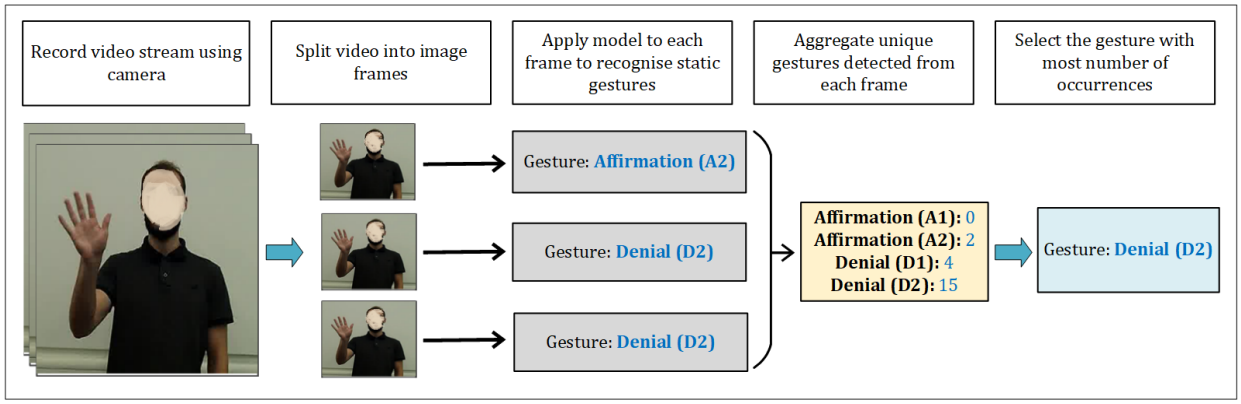

Fig. 3. Gesture recognition procedure.

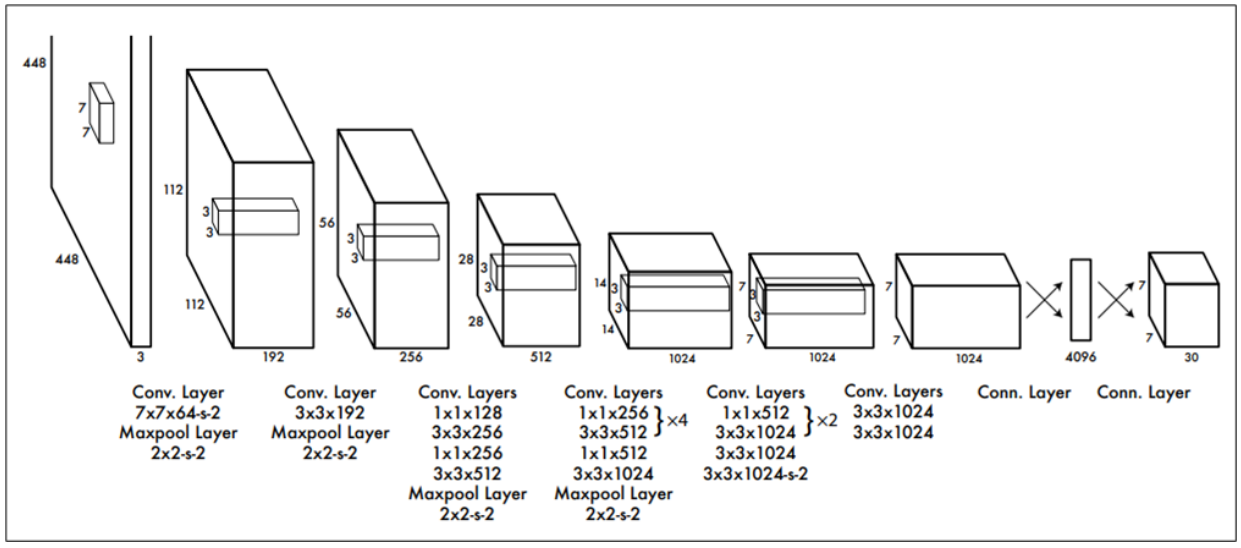

Fig. 4. YOLO architecture by Redmon et al. [21].

TABLE I
AN APPROXIMATE SUMMARY OF COLLECTED EXPERIMENTAL DATA.

Feature Description	Scenario 1	Scenario 2	Scenario 3
Number of participants	5	5	5
Number of Gestures	4	4	4
Length of each gesture	60 seconds	60 seconds	10 seconds X 10 = 10 videos
Length of all gestures per individual	240 seconds	240 seconds	40 videos
Length of all gestures for all individuals	1200 seconds	1200 seconds	200 videos
Extracted data per gesture	300 images	300 images	-
Extracted data of all gestures per individual	1200 images	1200 images	-
Extracted data of all gestures for all individuals	6000 images	6000 images	-
Training data size	2000 images	-	-
Testing data size	4000 images	6000 images	200 videos

position/orientation, and distance from the camera so that a generalised dataset can be obtained. The recorded video streams of the participants performing the gestures are then converted into image frames at the rate of 5 Frames Per Second (FPS). Each of the performed gesture generates approximately 300 image frames and a total of approximately $1,200$ frames for all the 4 gestures per individual. All the 5 participants combined generated a total of approximately 6000 images.

The data is split into a training and testing set. To achieve a generalised set, the split is carried out at an individual level i.e. the training and testing data are split for each participant separately before merging the whole data together. A split ratio of "1:2" is adopted for the training and testing respectively i.e. 33.33% of the data is used for training while the remaining data is used for testing. Overall, approximately 2000 images are used for training with over 500 image samples per gesture

Fig. 5. Sample output of the gesture recognition model.

SCENARIO 1		Predicted				
		Denial (D1)	Denial (D2)	Affirmation (A1)	Affirmation (A2)	No Gesture
Actual	Denial (D1)	901	0	35	0	51
	Denial (D2)	6	930	0	0	70
	Affirmation (A1)	5	0	840	0	161
	Affirmation (A2)	0	0	0	975	11
	No Gesture	0	0	0	0	0

Fig. 6. Result for Scenario 1.

SCENARIO 2		Predicted				
		Denial (D1)	Denial (D2)	Affirmation (A1)	Affirmation (A2)	No Gesture
Actual	Denial (D1)	764	161	78	42	426
	Denial (D2)	21	1030	36	130	255
	Affirmation (A1)	26	146	936	0	364
	Affirmation (A2)	0	31	0	1326	130
	No Gesture	0	0	0	0	0

Fig. 7. Result for Scenario 2.

SCENARIO 3		Predicted				
		Denial (D1)	Denial (D2)	Affirmation (A1)	Affirmation (A2)	No Gesture
Actual	Denial (D1)	50	0	0	0	0
	Denial (D2)	0	50	0	0	0
	Affirmation (A1)	0	0	50	0	0
	Affirmation (A2)	0	0	0	50	0
	No Gesture	0	0	0	0	0

Fig. 8. Result for Scenario 3.

while the remaining 4000 images are used for validation. While acknowledging the fact that the training set is usually larger than the test set, for this preliminary experiment, the small split ratio for the training data is chosen in order to minimise the labelling time and because the gestures across the data are expected to be the same. The labelled data is then used to train the YOLOv3 model. The aim is to ascertain if the trained model can identify the gestures in the test data.

In the second scenario termed as "Scenario 2", a separate dataset is collected from the 5 participants in a different environmental setting with relatively different lightning condition and background objects. Similar to "Scenario 1", the participants perform the 4 gestures with each gesture performed for 1 minute. This generates the same data size as in the previous scenario. The aim here is to test the robustness of the gesture detection model in a different environmental setting with different lightning and other physical conditions (i.e. environmental setting different from the training environment). This is important since in a real-life scenario, the robot intermediary may be utilised in an unknown environment.

For the third scenario (known as "Scenario 3"), the same 5 participants took part in the experiment in the same environmental setup as in "Scenario 1". This experiment aims to simulate a real-life scenario in which the gesture recognition approach is implemented on an assistive robot intermediary to communicate detected abnormalities and receive user's feedback in the form of hand gestures. The participants are asked to perform the gestures on instruction (i.e. the participants perform the gestures only when they are instructed). They are instructed to perform each gesture 10 times repeatedly. Each of the instructed gesture is recorded for 10 seconds. Images are extracted from the video at the rate of 5 FPS and the gesture recognition approach in Section III is applied. A total of 200 videos are obtained across the 5 participants. Table I contains an approximate summary of the collected data highlighting the number of participants, gestures, number of image frames obtained, as well as the size of the training and testing sets respectively.

The YOLOv3 confidence threshold reflecting the likelihood of an image containing one of the gestures is set to 0.5 across all the 3 scenarios i.e. a gesture is only predicted if the confidence value of the prediction is above 0.5. The threshold for the Non-Maximum Suppression responsible for the removal of duplicate prediction for the same class is also set to 0.5. In a situation where 2 or more unique gestures are predicted for the same image, the gesture with the highest confidence value is selected as the final prediction.

The model is implemented on a computer equipped with Intel Core i7 processor and NVIDIA GTX 1070 graphics card running Ubuntu 18.04. YOLO implementation in Darknet is used. Darknet which is a C based Neural Network (NN) framework is compiled with CUDA and OpenCV enabled. Over $20,000$ iterations are performed during the training and the weight generating the best result is selected for each experimental scenario.

IV. Experimental Result

Figure 5 shows a sample output of the gesture recognition model when applied to an image frame with the bounding box

TABLE II
PREDICTION RESULT FOR THE DIFFERENT GESTURES IN DIFFERENT SCENARIO.

	Scenario 1			**Scenario 2**			**Scenario 3**		
	Precision	Recall	F1 Score	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Denial (D1)	0.9879	0.9129	0.9489	0.9420	0.5194	0.6696	1.0000	1.0000	1.0000
Denial (D2)	1.0000	0.9245	0.9607	0.7529	0.6997	0.7254	1.0000	1.0000	1.0000
Affirmation (A1)	0.9600	0.8350	0.8931	0.8914	0.6359	0.7423	1.0000	1.0000	1.0000
Affirmation (A2)	1.0000	0.9888	0.9944	0.8852	0.8917	0.8884	1.0000	1.0000	1.0000

representing the identified gesture along with its confidence score. The results obtained from the experiment described in Scenario 1 is shown in Figure 6. The confusion matrix shows the different prediction results for the various gestures. It is worth mentioning that the training data is taken from the data collected in Scenario 1. The "No Gesture" entries in the confusion matrix represent a situation where no gestures are detected in the validation set.

Figure 7 shows the results obtained when the model is validated with the data collected in Scenario 2. It can be seen that the performance is significantly lower than that of Scenario 1. This may be due to the sampling of the training data from Scenario 1 only, resulting in overfitting because of the similarity in the environmental setup. The number of non detected cases is significantly large, indicating that the model is unable to predict a large portion of the data. Incorporating data collected from Scenario 2 into the training set may improve the model performance and reduce any possible bias.

The result for Scenario 3 shown in Figure 8 indicates that all the model's predictions are all correct. As mentioned earlier, Scenario 3 experiment is carried out in the same environment as Scenario 1. The excellent result obtained may be due to the similarity between the 2 scenarios since the model is trained on data collected in Scenario 1. Furthermore, the gesture recognition approach described in Section III aggregates the model predictions and selects the gesture with the highest number of occurrence as the final prediction. This approach is not likely to make incorrect predictions because even if some of the data are misclassified, the number of correctly classified data is likely going to be higher. In Table II, the Precision, Recall and F1 score of the model for all the experimental scenarios are summarised.

To test for generalisation of the model, a K-Fold Cross Validation (CV) is performed with $K = 5$. Since there are 5 participants for the experiment, each of the participant's data is used as 1 fold for the CV (i.e. data for 4 participants are used for the training while that of the 5th participant is used for testing and vice versa). Data collected from the experiment in Scenario 1 is used for the CV. The average accuracy of the 5-Fold CV is calculated and tabulated in Table III along with the overall accuracy of the other experimental scenarios.

TABLE III
OVERALL RECOGNITION ACCURACY.

Experimental Scenario	**Accuracy**
Scenario 1	0.9149
Scenario 1 (5-Fold CV)	0.8428
Scenario 2	0.6872
Scenario 3	1.0000

V. CONCLUSION AND FUTURE WORK

In this paper, a gesture recognition approach is proposed for utilisation in an assistive robot intermediary for abnormality detection in ADL. Human activities classified as abnormal by the anomaly detector will be communicated to humans through the intermediary for confirmation. Human response in the form of hand gestures will be fed back into the model to improve the model's accuracy. A gesture recognition model is implemented for the detection of 4 hand gestures corresponding to affirmations and denials in 3 different experimental scenarios. The obtained results show that the proposed approach achieved good predictive accuracy and has the potential of been utilised in the intermediary and other related purposes.

The experiments are conducted in a controlled lab environment with middle-aged individuals under similar lightning and other physical conditions. This may not be the case in a real home deployment. Extensive experiments will be carried out under different conditions with data collected from older adults to ascertain the effectiveness of the proposed gesture recognition approach. Depth sensor capable of generating RGB-Depth data may be considered over 2D camera due to its resilience to physical constraints such as variability in background and lighting condition. In the current approach, we focus only on the detection of static gestures for a single individual. Dynamic gestures will be considered as well as approaches for dealing with multiple subjects in the robot's line of sight. Furthermore, future work will involve a comparison of the proposed approach with baseline models, exploring avenue for the fusion of the gesture recognition model with other interaction modalities (e.g. speech and touch input), as well as the incorporation of the gesture recogniser and an assistive robot into the overall anomaly detection system.

REFERENCES

[1] S. Chernbumroong, S. Cang, A. Atkins, and H. Yu, "Elderly activities recognition and classification for applications in assisted living," *Expert Systems with Applications*, vol. 40, no. 5, pp. 1662 – 1674, 2013.

[2] D. Arifoglu and A. Bouchachia, "Detection of abnormal behaviour for dementia sufferers using Convolutional Neural Networks," *Artificial Intelligence in Medicine*, vol. 94, pp. 88–95, 2019.

[3] A. A. Aramendi, A. Weakley, A. A. Goenaga, M. Schmitter-Edgecombe, and D. J. Cook, "Automatic assessment of functional health decline in older adults based on smart home data," *Journal of Biomedical Informatics*, vol. 81, pp. 119 – 130, 2018.

[4] S. W. Yahaya, A. Lotfi, and M. Mahmud, "A consensus novelty detection ensemble approach for anomaly detection in activities of daily living," *Applied Soft Computing*, vol. 83, p. 105613, 2019.

[5] A. Lotfi, C. Langensiepen, S. M. Mahmoud, and M. J. Akhlaghinia, "Smart homes for the elderly dementia sufferers: identification and prediction of abnormal behaviour," *Journal of Ambient Intelligence and Humanized Computing*, vol. 3, no. 3, pp. 205–218, Sep 2012.

[6] E. Hoque, R. F. Dickerson, S. M. Preum, M. Hanson, A. Barth, and J. A. Stankovic, "Holmes: A comprehensive anomaly detection system for daily in-home activities," in *International Conference on Distributed Computing in Sensor Systems*. Fortaleza, Brazil: IEEE, June 2015, pp. 40–51.

[7] S. W. Yahaya, A. Lotfi, and M. Mahmud, "A framework for anomaly detection in activities of daily living using an assistive robot," in *2nd UK-RAS Robotics and Automation Conference*, 01 2019, pp. 131–134.

[8] J.-J. Cabibihan, H. Javed, M. Ang, and S. M. Aljunied, "Why robots? a survey on the roles and benefits of social robots in the therapy of children with autism," *International Journal of Social Robotics*, vol. 5, no. 4, pp. 593–618, 2013.

[9] J. Wainer, D. J. Feil-seifer, D. A. Shell, and M. J. Mataric, "The role of physical embodiment in human-robot interaction," in *15th IEEE International Symposium on Robot and Human Interactive Communication*, 2006, pp. 117–122.

[10] H. . Gross, C. Schroeter, S. Mueller, M. Volkhardt, E. Einhorn, A. Bley, T. Langner, C. Martin, and M. Merten, "I'll keep an eye on you: Home robot companion for elderly people with cognitive impairment," in *2011 IEEE International Conference on Systems, Man, and Cybernetics*, 2011, pp. 2481–2488.

[11] A. Lotfi, C. Langensiepen, and S. W. Yahaya, "Socially assistive robotics: Robot exercise trainer for older adults," *Technologies*, vol. 6, no. 1, 2018.

[12] L. Cañamero and M. Lewis, "Making new "new ai" friends: Designing a social robot for diabetic children from an embodied ai perspective," *International Journal of Social Robotics*, vol. 8, no. 4, pp. 523–537, 2016.

[13] L. Wood, B. Robins, G. Lakatos, D. S. Syrdal, A. Zaraki, and K. Dautenhahn, "Utilising humanoid robots to assist children with autism learn about visual perspective taking," in *1st UK-RAS Conference on Robotics and Autonomous Systems*, 12 2017.

[14] D. H. Neiva and C. Zanchettin, "Gesture recognition: A review focusing on sign language in a mobile context," *Expert Systems with Applications*, vol. 103, pp. 159 – 183, 2018.

[15] R. C. Luo, Y. C. Wu, and P. H. Lin, "Multimodal information fusion for human-robot interaction," in *2015 IEEE 10th Jubilee International Symposium on Applied Computational Intelligence and Informatics*, 2015, pp. 535–540.

[16] J. Nagi, F. Ducatelle, G. A. Di Caro, D. Cirean, U. Meier, A. Giusti, F. Nagi, J. Schmidhuber, and L. M. Gambardella, "Max-pooling convolutional neural networks for vision-based hand gesture recognition," in *2011 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, 2011, pp. 342–347.

[17] C.-B. Park and S.-W. Lee, "Real-time 3d pointing gesture recognition for mobile robots with cascade hmm and particle filter," *Image and Vision Computing*, vol. 29, no. 1, pp. 51 – 63, 2011.

[18] X. Yingxin, L. Jinghua, W. Lichun, and K. Dehui, "A robust hand gesture recognition method via convolutional neural network," in *2016 6th International Conference on Digital Home (ICDH)*, 2016, pp. 64–67.

[19] C. Feichtenhofer, A. Pinz, and A. Zisserman, "Convolutional two-stream network fusion for video action recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[20] Z. Hu, Y. Hu, J. Liu, B. Wu, D. Han, and T. Kurfess, "3d separable convolutional neural network for dynamic hand gesture recognition," *Neurocomputing*, vol. 318, pp. 151 – 161, 2018.

[21] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: Unified, real-time object detection," in *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 779–788.