Optimised ML-based System Model for Adult-Child Actions Recognition

Muhammad Alhammami[1*], Samir Marwan Hammami[2], Chee-Pun Ooi[3] and Wooi-Haw Tan[4]
[1, 3, 4] Faculty of Engineering, Multimedia University
Cyberjaya, 63000 – Malaysia
[2] Department of Management Information, Dhofar University.
Salalah 211, 2509 - Oman
[1] [email: dr.mhammami@outlook.com]
[2] [email: samir@du.edu.om]
[3] [email: cpooi@mmu.edu.my]
[4] [email: twhaw@mmu.edu.my]
*Corresponding author: Muhammad Alhammami

ReceivedOctober 21, 2017; revised July 15, 2018; accepted September 12, 2018; published February 28, 2019

Abstract

Many critical applications require accurate real-time human action recognition. However, there are many hurdles associated with capturing and pre-processing image data, calculating features, and classification because they consume significant resources for both storage and computation. To circumvent these hurdles, this paper presents a recognition machine learning (ML) based system model which uses reduced data structure features by projecting real 3D skeleton modality on virtual 2D space. The MMU VAAC dataset is used to test the proposed ML model. The results show a high accuracy rate of 97.88% which is only slightly lower than the accuracy when using the original 3D modality-based features but with a 75% reduction ratio from using RGB modality. These results motivate implementing the proposed recognition model on an embedded system platform in the future.

Keywords: Human action recognition, 2D Skeleton features, 3D Projection, Reduced data structure, Compound features selection method

1. Introduction

Human action recognition is a developing technology requiring complex operations to extract useful information from image sequences or videos. A typical human action recognition system consists of standard steps that start with image acquisition, image pre-processing, features extraction and classification. The bigger the image data, the more complex the underlying operations and the slower the performance. Increasing the accuracy rates regardless of the complexity of the data structures and algorithms is not suitable when planning to implement the recognition system model as an embedded system. One viable solution in optimising the recognition model is to improve the quality of the modalities and features. The usage of skeletal data obtained from depth sensors is promising in extracting features for action recognition. However, using all the skeleton data as features will significantly increase the complexity of the system and dramatically decrease the performance.

This paper presents a machine learning based system model for recognising Adult → Child actions. These activities have not been previously investigated despite its many critical applications such as detecting child abuse detection. It is a new area of research due to its broad application in many areas such as management information systems. The proposed model uses 25 features based on skeletal joints modality after projecting it from 3D real space on 2D planar space. We decrease the dimensionality of the features further by applying two cascaded algorithms during the feature selection step. The first algorithm will be a scheme-independent algorithm while the second will be a scheme-dependent algorithm.

Section 2 reviews previous related work in the field of human action recognition. Section 3 presents the methodology used during this work, while section 4 presents the experimental evaluations of the recognition model.

2. Related Work

Human activities take place in 3D space. This renders depth information essential to successful recognition [1]. A skeletal modality is an important approach used commonly to distinguish human actions based on human skeletal movements. Skeletal joints encode the 3D positions of human joints per frame in real time. Comparing with RGB data based skeleton structures, the depth information makes the modelling more feasible and stable. Several algorithms were proposed and applied to model the skeleton from the depth data [2-4]. The basic idea underlying these methods is to segment the depth data of the human body into multiple parts with dense probabilistic labelling. Over the past few years, significant progress in pose estimation has been reported [5-8]. Hence, many types of research in action recognition were done by using the pose-based methods. Early works in recognising human activity which focused on tracking body parts and classifying the motion of joints suffered from inaccurate tracking [9-11]. More recent works assumed that poses are readily available from an independent tracker [12] or parts labelling [13], or they used viewpoint variant 2D pose estimation [14]. An example of a hardware tracker is the Microsoft Kinect sensor. The first generation of Kinect provides 20 joints for each video frame, while the second generation Kinect allows up to 25 joints.

2.1 Skeletal Based Datasets of Human-Human Actions

There are many datasets of human-human (two persons) and multi-human (more than two persons) interactions. [15] and [16] contain datasets with eight types of interactions: approaching, departing, pushing, kicking, punching, exchanging objects, hugging, and shaking hands. The LIRIS daily-life human activities dataset is recorded using a Kinect sensor fixed on a robot capturing human-human and human-object interactions [17]. The G3di dataset is a multi-player gaming scenarios dataset [18]. There is also a dataset about office activities [19]. The human-human interactive part of the M2I dataset includes handshaking, high-fives, bowing, walking, crossing, waiting, chatting, hugging, and boxing [20].

This research uses the MMU VAAC dataset containing adult → child actions. The recorded actions include kicking, punching, throwing, shoving, strangling and slapping, touching, hugging, carrying, laying down, departing, and approaching [21].

2.2 Skeletal Based Features

Skeletal features can be sorted into space-time and sequential-based features [22]. Space-time features are extracted by representing an action sequence as a space-time volume, while sequential features are based on frame-level or subsequence-level information considering any human action as a sequence of temporal observations.

2.2.3 Space-Time Features

The features are usually extracted from the skeletal joints to encode temporal motion information. [23] developed "Eigenjoints" features from skeleton data of action sequences. An "Eigenjoint" engages the different joint positions to express human actions. "Eigenjoints" are extracted by applying PCA which encodes the most valuable information about the actors' poses for their action recognition.

Ref. [24] utilised point cloud and skeletal information. Ref. [25] proposed a new discriminative dictionary learning algorithm (DL-GSGC) that uses both geometry constraints and group sparsity to describe skeletal features better. The previous skeleton based space-time features use information purely about the joints like locations or angles to express human actions. Ref. [26] used translations and rotations in 3D space to model the 3D geometric relationships among various body parts. Ref. [15] also used geometric relational features based on the Euclidean distance between each pair of joints.

2.2.4 Sequential Features

Sequential features are typically extracted from skeletal joints within frames or to model temporal dynamics of actions. Skeleton-based sequential features produce low-latency responses allowing the recognition of any action even before it ends. Ref. [27] presented a Histogram of 3D Joint Locations (HOJ3D) feature which uses modified spherical coordinates. HOJ3D can encode primarily spatial occupancy information about the skeleton hip centre. Ref. [28] proposed that Structured Streaming Skeletons (SSS) can handle intra-class variations including viewpoint, anthropometry, execution rate, and personal style. To cope with unsegmented action sequences, Ref. [29] proposed a non-parametric moving pose (MP) descriptor for low-latency human action recognition. Ref. [30] proposed a dynamic hierarchical framework based on high-level skeletal joints features to segment and recognise actions simultaneously. They made feature extraction from skeleton data an implicit approach using deep belief networks.

3. The Recognition Model

This section presents the methods adopted for analysing and developing the proposed model. **Fig. 1** illustrates the research process. While creating the dataset has been discussed in [21], skeleton tracking will be done using the depth sensor. The research determines the suitable features and classifier for this model is performed once offline during the features extraction and classification phases. After that, any new data will be passed to the resulted model to predict the action in each frame.

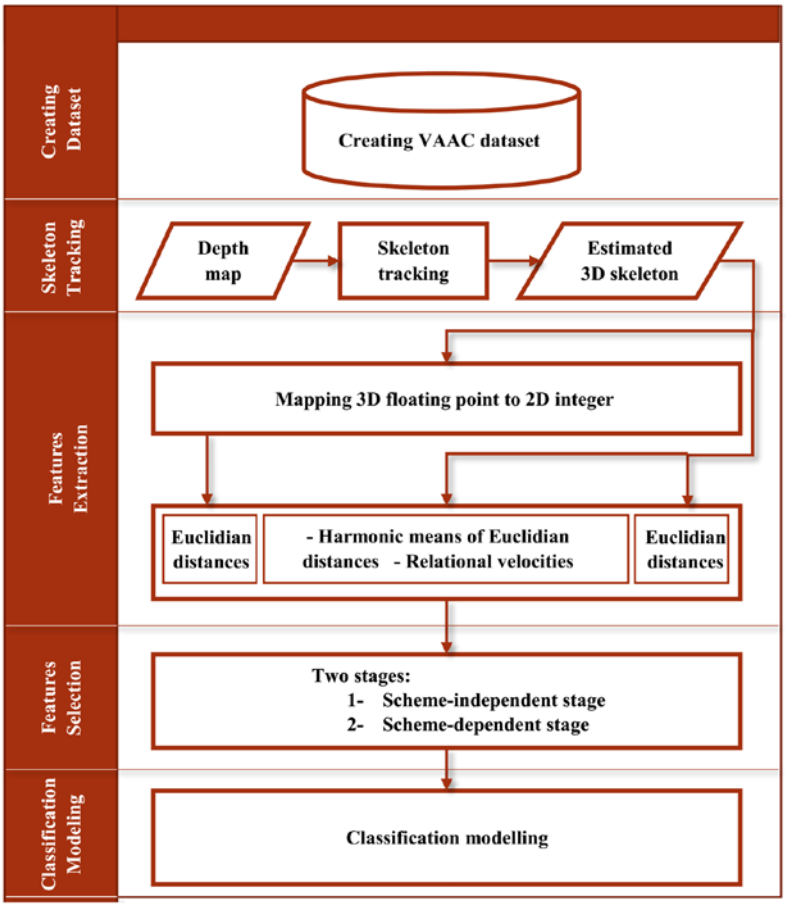

Fig. 1. Stages for developing the recognition model

3.1 Choosing the Modality

The research uses skeleton modality because it has the following attributes:

1) Robust in a noisy environment.
2) Viewpoint invariant model for 3D human pose extraction.
3) Not affected by variations in people's weight.
4) Not affected by changes in lighting condition.
5) Can be extracted in real time by depth sensors like Kinect without the need of extra hardware resources.
6) Requires a smaller data structure in comparison to RGB data according to Eq. (1) and Eq. (2).

$$Framesize = 640 \times 480 \times 3RGB \times 8bit = 7200Kbit \quad (1)$$

Equation (1) shows that for each frame having a resolution of (640 × 480) pixels needs to manage 7.2Mbit where each colour channel usually needs 8 bits to represent the values from 1 up to 255. The skeleton modality requires 3.75Kbit as demonstrated in Eq. (2) where each joint in the 3D real space is represented in floating point format, and two people have 20 joints for each.

$$Framesize = 2persons \times 20joints \times 3coordinate \times 32bit = 3.75Kbit \quad (2)$$

3.2 Projecting Joints Coordinates on Virtual 2D Planar Space

It is beneficial to reduce the bit-width of the skeleton modality further. Fortunately, this is possible by projecting the real 3D floating point coordinates on a 2D virtual planar. The data size of each frame is 0.9375Kbit if the 2D planar space has 680x480 pixels according to Eq. (3). This involves reducing 99.98% of the size when using RGB modality and 75% when using the 3D real coordinates in Eq. (2). The number of bits in Eq. (3) is 12 bits because the coordinates are represented in an integer with a maximum value of 640 which is the resolution of the 2D planar.

$$Framesize = 2persons \times 20joints \times 2coordinate \times 12bit = 0.9375Kbit. \quad (3)$$

The projection method of the 3D real space joints on a 2D planar space coordinates is used to calculate one of the proposed features. Considering **Fig. 2**, A(x_a, y_a, z_a) is a joint in the real 3D space. We need to calculate its projection point A'(m,n) on the virtual 2D planar coordinates. The coordinates of the four corners of the 2D planar (S1, S2, S3, S4), and hence the distance to Kinect, are chosen depending on the required resolution. To achieve this, the following steps are followed:

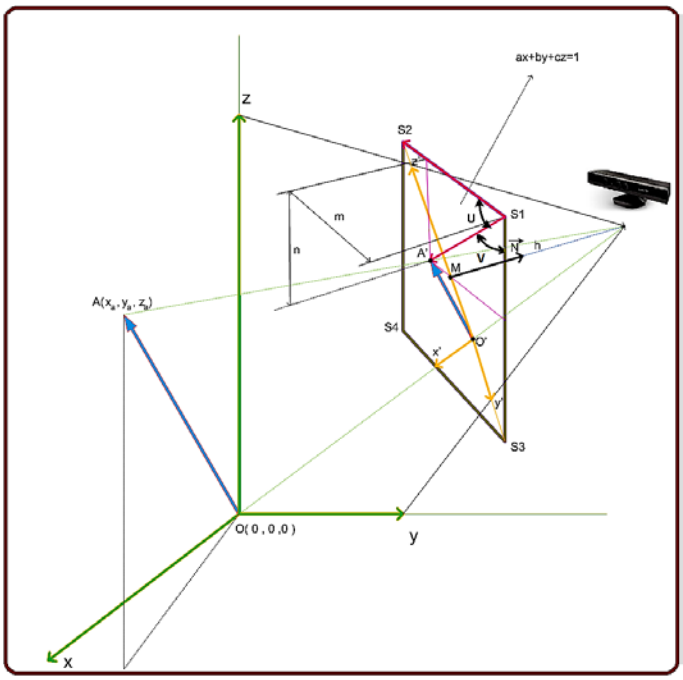

Fig. 2. Projecting real 3D point on 2D virtual planar space

- Let the left top point of the 2D planar $S_1(x_1, y_1, z_1)$, and $S_2(x_2, y_2, z_2)$ be at the right top point of 2D planar in the plane. The width (W) of 2D planar satisfies:

$$W = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \tag{4}$$

For a straight view, select $z_1=z_2$ for

$$W = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \tag{5}$$

- Let the left bottom point of the 2D planar in the plane be $S_3(x_3, y_3, z_3)$. The height (H) of the 2D planar satisfies:

$$H = \sqrt{(x_3 - x_1)^2 + (y_3 - y_1)^2 + (z_3 - z_1)^2} \tag{6}$$

- Since the 2D planar has a rectangular shape, it satisfies:

$$\vec{S_1S_2}.\vec{S_1S_3} = 0$$

$$\Rightarrow (x_2 - x_1)(x_3 - x_1) + (y_2 - y_1)(y_3 - y_1) + (z_2 - z_1)(z_3 - z_1) = 0 \tag{7}$$

- For a straight view, $x_1=x_3$ and $y_1=y_3$ thus:

$$H = z_1 - z_3 \tag{8}$$

- Let $M(x_0, y_0, z_0)$ be the centre point of the 2D planar:

$$M(x_0, y_0, z_0) = (\frac{x_2 + x_3}{2}, \frac{y_2 + y_3}{2}, \frac{z_2 + z_3}{2}) \tag{9}$$

- Let h be the distance between the Kinect sensor $C(x_c, y_c, z_c)$ and the virtual normalised 2D planar which has a planar equation of the shape:

$$ax + by + cz = 1 \tag{10}$$

- The points S_1, S_2, and $M \in P$ are in Eq. (10):

$$\begin{aligned} S_1 &: ax_1 + by_1 + cz_1 = 1 \\ S_2 &: ax_2 + by_2 + cz_2 = 1 \\ M &: ax_0 + by_0 + cz_0 = 1 \end{aligned} \tag{11}$$

- Solving a, b, c and finding the normalisation vector (N) gives:

$$\vec{N} = (a_n, b_n, c_n) = (\frac{a}{\sqrt{a^2 + b^2 + c^2}}, \frac{b}{\sqrt{a^2 + b^2 + c^2}}, \frac{c}{\sqrt{a^2 + b^2 + c^2}}) \tag{12}$$

where:

$$C(x_c, y_c, z_c) = (x_0 + ha_n, y_0 + hb_n, z_0 + hc_n) \tag{13}$$

- To find $A'(x'_a, y'_a, z'_a)$, the projection of any point $A(x_a, y_a, z_a)$ in the 3D real space on the 2D planar, let the line between point $C(x_c, y_c, z_c)$ and point A be:

$$\frac{x - x_a}{x_c - x_a} = \frac{y - y_a}{y_c - y_a} = \frac{z - z_a}{z_c - z_a} = k \tag{14}$$

- Putting x, y, z into the plane equation (10) and getting an equation depends on k. To solve k, we get A' from Eq. 14.
- To find the 2D planar angles (u and v):

$$\cos u = \frac{\vec{S_1S_2}.\vec{S_1A'}}{|\vec{S_1S_2}||\vec{S_1A'}|} = \frac{(x_2 - x_1)(x'_a - x_1) + (y_2 - y_1)(y'_a - y_1) + (z_2 - z_1)(z'_a - z_1)}{W\sqrt{(x'_a - x_1)^2 + (y'_a - y_1)^2 + (z'_a - z_1)^2}} \tag{15}$$

$$\cos v = \frac{\vec{S_1S_3}.\vec{S_1A'}}{|\vec{S_1S_3}||\vec{S_1A'}|} = \frac{(x_3 - x_1)(x'_a - x_1) + (y_3 - y_1)(y'_a - y_1) + (z_3 - z_1)(z'_a - z_1)}{H\sqrt{(x'_a - x_1)^2 + (y'_a - y_1)^2 + (z'_a - z_1)^2}}$$

- If cos(u) > 0 and cos(v) > 0 then A' is in the 2D planar. Otherwise, point A' is outside of the 2D planar, and we cannot draw A' in the 2D planar. Such are the points behind the Kinect sensor, or out of its view, and hence are not important points in this application.
- Finding the new projected coordinates (m, n) in the 2D space when cos(u) > 0 and cos(v) > 0:

$$m = \sqrt{(x'_a - x_1)^2 + (y'_a - y_1)^2 + (z'_a - z_1)^2}\cos u$$
$$n = \sqrt{(x'_a - x_1)^2 + (y'_a - y_1)^2 + (z'_a - z_1)^2}\sin u \quad (16)$$

- We ignore the fractional part in m and n to get integer values. If m>W and n > H_n > H, then it is not possible to project the joint on this 2D planar.

3.3 Features Extraction

In this step, relational features are used for encoding the pose information. Using these features helps describe the geometric relationships between joints in a single frame or a short sequence of frames. This is because the relational features are more robust to spatial variations than the poses themselves, and semantically, similar motions belonging to the same action are not necessarily numerically similar [31].

This research experiments on three types of features to evaluate the proposed features structure (i.e., Type III) by comparing it to two features used in previous literature:

- Type I: Harmonic means of relational Euclidean distances and velocities in sliding windows in the 3D real space. Harmonic mean is applied on geometric relational distances to deal with irrelevant frames in sliding windows, and the relational velocities give a sense of the speed of movements [32].
- Type II: Relational Euclidean distances in each frame in the 3D real space.
- Type III: Relational Euclidean distances in each frame in the projected 2D planar space.

Type III features are the proposed optimised features with the smallest data structure which are benchmarked using type I and type II features. It is important to mention that the determination of having an adult and a child in the scene is taken merely by measuring the height of the subjects, so we assume later that all scenes consist of an adult and child.

3.3.1 Euclidean Distance Features

Euclidean distance (d) between any two points (p_1, p_2) in 3D space is given by Eq. 17:

$$\| d(p_1(x_1, y_1, z_1), p_2(x_2, y_2, z_2))\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2} \quad (17)$$

Moreover, in 2D space the Euclidean distance is given by Eq. 18:

$$\| d(p_1(x_1, y_1), p_2(x_2, y_2))\| = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2} \quad (18)$$

3.3.2 Harmonic Mean of Relational Euclidean Distances Features

For this type of feature, we consider sliding windows of W frames. The value of W is determined by trials during the research process to reach the best accuracy. The harmonic mean of each Euclidean distance is calculated for each sliding window. Since the harmonic mean of a list of numbers tends to tend strongly towards the least elements of the list, it has the advantage of mitigating the impact of outliers with large values. So, using the harmonic mean

helps address the irrelevant frames around the main action and mitigate the impact of odd and wrong values extracted by the depth device. The harmonic mean of Euclidean distances d_t between any two joints J_x and J_y in a window with W frames is Eq. (19):

$$H_{J_x,J_y} = \frac{W}{\sum_1^w \frac{1}{d_t}} \tag{19}$$

3.3.2 Relational Velocities Features

The relational velocities features (V) could offer a significant clue about how much action is violent. The relational velocity of any two joints in a window is obtained via Eq. (20):

$$V_{J_x,\ J_y} = \frac{\max d_t - \min d_t}{W_{max} - W_{min}} \tag{20}$$

where J_x and J_y are two joints from the adult or the child, i and j are indexes of joints [1...20], t is an index of frames in a window [1...W], and d_t is a Euclidean distance between any two joints in one frame. $W_{min,\ max}$ is the frame number in [1...W], where the minimum and maximum Euclidean distances between two joints were respectively found in each window.

3.4 Features Selection

Features selection is essential to reduce the needed resources in any future implementation stage further. This is done by choosing the most relevant features to classes by excluding redundant, noisy, and irrelevant features. It also reduces the dimensionality of the data to allow better performance. There are two criteria kept true while features selection is performed. Firstly, the classification accuracy does not significantly decrease. Secondly, the resulting class distribution, given only by the values for the selected features, is as close as possible to the original class distribution given by all features.

Two stages of feature selection have been applied. The first stage is scheme-independent and the second stage is scheme-dependent. In the first stage, all correlated features will be eliminated. However, it may give a significant number of features. Hence, the second stage is vital to rank the resulting features individually. It is important to notice that, alone, the second stage is unable to remove correlated or redundant features. Thus, this compound approach is needed to reduce the dimensionality of features as much as possible.

3.4.1 Scheme-independent

Correlation-based Feature Selection (CFS) algorithm is firstly applied using the BestFirst searching method. The BestFirst algorithm searches the space of features subsets by greedy hill climbing augmented with a backtracking facility to evaluate features and select the subset (S) of features highly correlated with the class but not strongly correlated with each other. The heuristic merit for this subset of features is given by [33] in Eq. (21):

$$Merit_S = \frac{k\overline{r_{ca}}}{\sqrt{k + k(k-1)\overline{r_{aa}}}} \tag{21}$$

where r_{ca} is the average feature-class correlation, k is the number of features, r_{aa} is the average feature-feature correlation. This heuristic value reflects the usefulness of distinctive features for anticipating the class label forth with the measure of inter-correlation among them.

3.4.2 Scheme-dependent

Secondly, we employed the scheme-ranking of the output features selected by CFS by measuring the gain ratio on the class, Eq. (22).

$$GainR(Class, Attribute) = \frac{H(Class) - H(Class|Attribute)}{H(Attribute)} \quad (22)$$

Where H is the entropy value, we draw the learning curves of all possible number of features extracted to determine the best number of features to be used in the system model. The reason behind this is that the scheme-ranking approach does not give the required number of features explicitly.

3.5 Classification

The classification phase aims to find the best model which gives the highest possible recognition accuracy rate. Finding the right algorithm is partly trial and error by evaluating the most algorithms mentioned in the literature of vision-based human action recognition. The influence of each key parameter in each algorithm is investigated for the highest accuracy. For each algorithm, N folds cross-validation technique is performed by repeating each experiment 10 times. The benefit of this procedure is to increase the reliability of the verification results and to check the model against over-fitting. All necessary curves and tables for best comparison between classifiers are extracted.

4. Experimental Results and Analysis

We used a JAVA-based platform for experimenting and evaluating the proposed model.

4.1 Dataset Validation

The MMU VAAC dataset is used for evaluations. Firstly, we had to validate the dataset. For this purpose, we drew the learning curves of the action name class as a function of the dataset size to ensure that the dataset is big enough and consistent to continue evaluation. Many classification algorithms were tested during the work, but we focused on three classification algorithms which gave the highest three recognition rates: k-NN, Random Forest, and SVM. Figures 3, 4, and 5 show that the learning curves are incremental functions to the used amount of the input data. It is also evident that using any of the three types of features, K-NN gives the best accuracy, then random forest and lastly SVM. We notice that the three classifiers need approximately 80% of the dataset to reach the maximum possible accuracy rate. Hence, a 5-fold cross-validation technique is used in the remainder of this research.

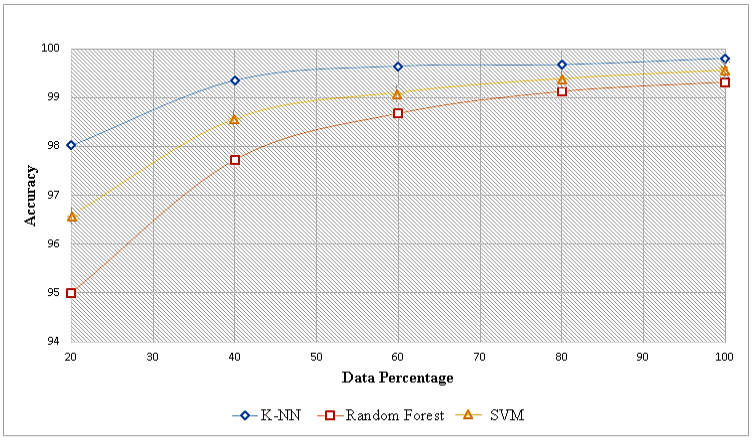

Fig. 3. Comparison of learning curves of k-NN, Random Forest, and SVM classifiers applied on type I features

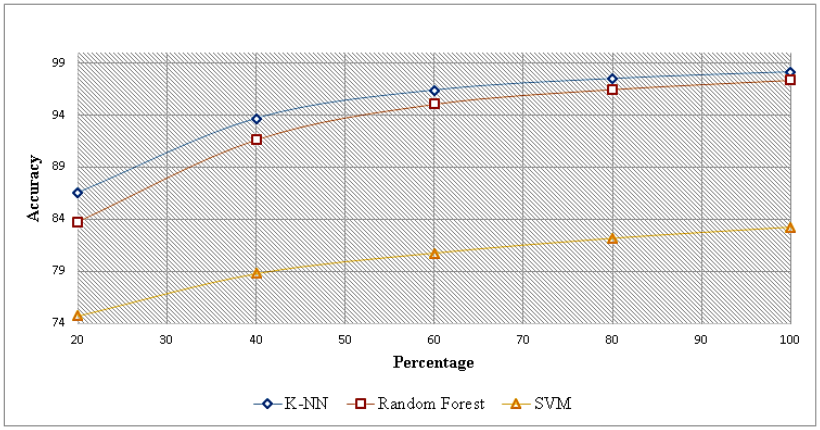

Fig. 4. Comparison of learning curves of k-NN, Random Forest, and SVM classifiers applied on type II features

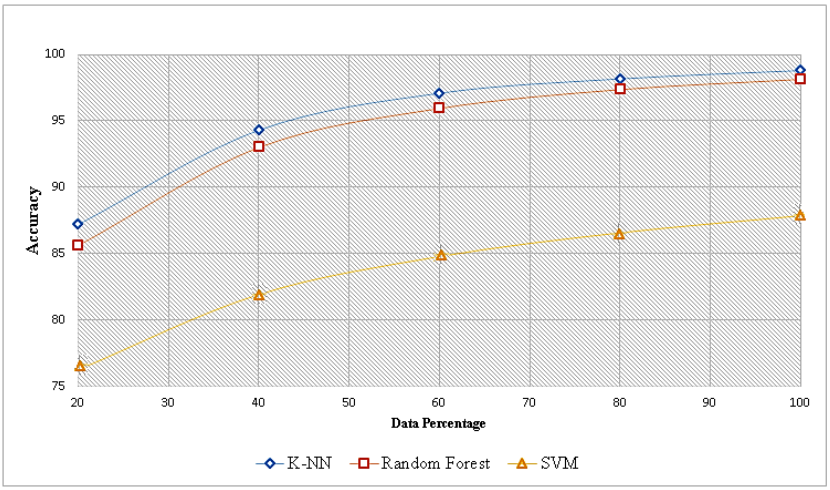

Fig. 5. Comparison of learning curves of k-NN, Random Forest, and SVM classifiers applied on type III features

4.2 Features Selection

Features selection is then done to decrease the dimensionality of all types of features for realistic comparisons. Both scheme-independent and scheme-dependent stages were applied. The first selection stage uses the CFS algorithm which selects the subset of features which are highly correlated with the class but uncorrelated with each other. The number of features of CFS algorithm subsets for each type of feature is shown in **Table 1**.

Table 1. Number of features in the first selection stage CFS algorithm

Type	Number of Features
Type I: Harmonic Means of Relational Distances and Relational Velocities in 3D Real Space	230
Type II: Relational Distances in 3D Real Space	80
Type III: Relational Distances in 2D Plane Space	74

In the second stage of the features selection phase, we applied a learning scheme-based ranking to determine the minimum sufficient number of features. To achieve this, the learning curves as functions of the number of features are analysed for the three classifiers (k-NN, Random Forest, and SVM). As shown in **Fig. 6**, the scheme-dependent rankings using information gain measure shows that using type II and III features almost reached its maximum accuracy rate using 25 features while using type I features needs 80 features. Hence, we will use the top 25 features of type III since it has a smaller data structure than type II. These features are shown in **Table 2**. Using the features in **Table 2** and the 1-NN classifier achieves an accuracy rate of 97.88%. The corresponding confusion matrix in **Fig. 7** shows excellent measures. However, the most missed predicted actions are: departing, hugging and approaching a child as they have the least recognition rates (93%, 94%, 95%) respectively as shown in **Fig. 7**.

5. Conclusion and Future Work

This paper presented an ML-based system model for recognising adult → child actions using optimised joints based features projected on 2D virtual planar coordinates instead of 3D real space. The primary goal was to decrease the data bits needed for the input of the system model, and eventually to reduce the resources required in any future embedded real-time implementation. One significant contribution of this work is analysing adult → child actions for the first-time. This paper investigated using a double stage feature extraction which achieved a high recognition accuracy of 97.88%. The enclosed results encouraged us to start implementing the above ML-based model, and we have already set up an embedded based framework which consists of a Kinect sensor, a CPU-based system to interface the Kinect device, and an FPGA to accelerate the computation of features and the classification as shown in **Fig. 8**.

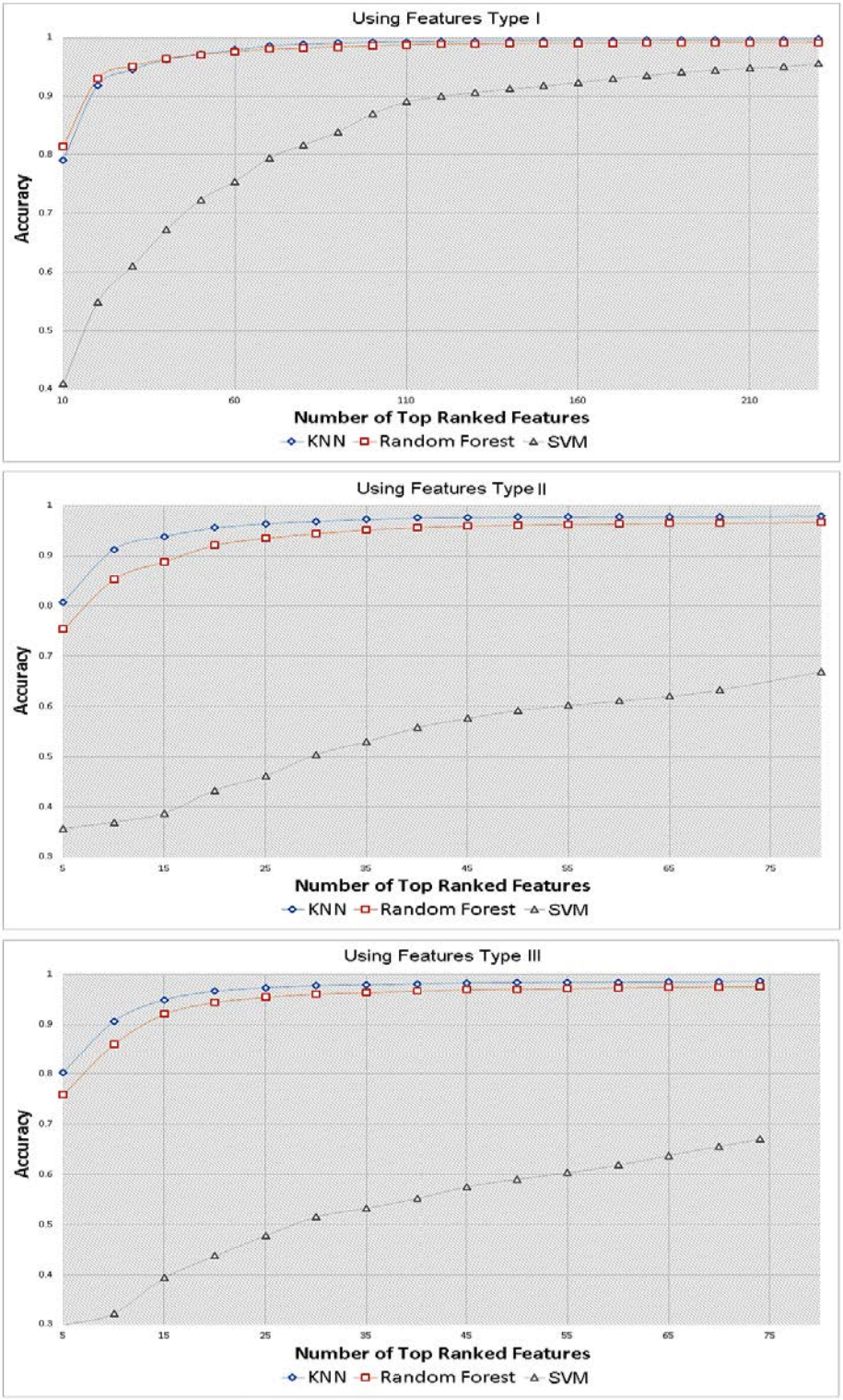

Fig. 6. Learning curves as functions of the number of tops ranked features in the second stage (scheme-dependent ranking) for K-NN, Random Forest, and SVM classifiers applied to all types of features

Table 5. The final selected 25 features with their gain ratios

Gain Ratio	Euclidean Distances Between Joints
0.15	child's shoulders' centre ←→ adult's shoulders' centre
0.14	child's head ←→adult's head
0.14	child's left shoulder ←→adult's left shoulder
0.18	child's left shoulder ←→adult's right shoulder
0.14	child's head ←→adult's shoulders' centre
0.13	child's right shoulder ←→adult's right shoulder
0.13	child's shoulders' centre ←→adult's right shoulder
0.12	adult's head ←→adult's left foot
0.12	child's right elbow ←→adult's right elbow
0.12	child's left elbow ←→adult's spine
0.12	child's left shoulder ←→adult's right elbow
0.12	child's left elbow ←→adult's left elbow
0.12	child's right shoulder ←→adult's shoulders' centre
0.12	child's left hip ←→adult's right ankle
0.12	child's left knee ←→adult's right ankle
0.12	adult's head ←→adult's left ankle
0.12	adult's shoulders' centre ←→adult's right foot
0.11	child's right knee ←→adult's right ankle
0.11	adult's head ←→ adult's left knee
0.11	child's left knee ←→adult's right foot
0.11	child's left elbow ←→ adult's left wrist
0.11	child's head ←→ adult's left hand
0.11	child's left shoulder ←→ adult's right wrist
0.11	child's left ankle ←→ adult's left ankle
0.11	child's right shoulder ←→ child's right foot

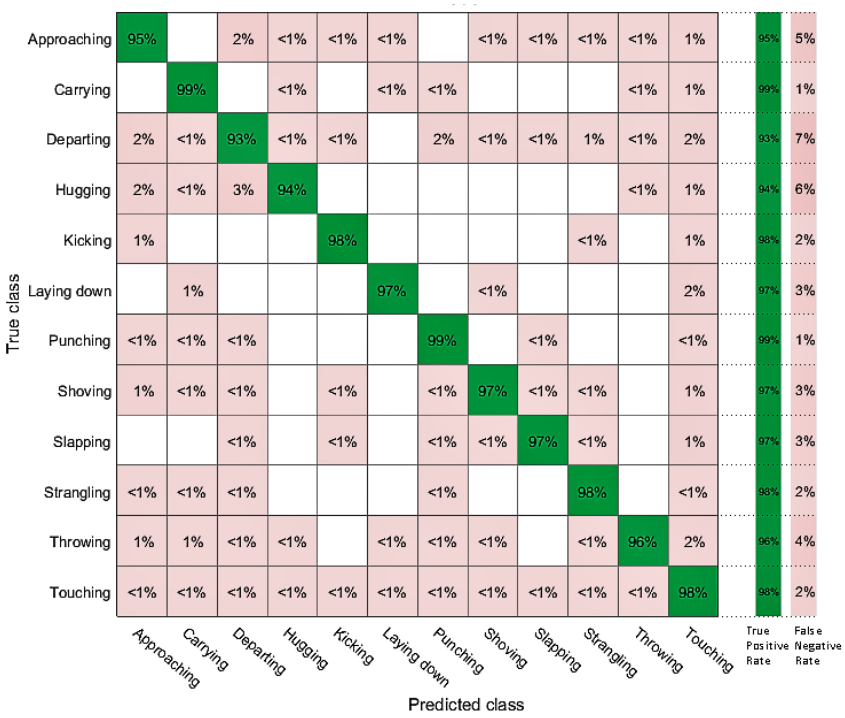

Fig. 7. Confusion Matrix and True Positive rate and False Negative rate

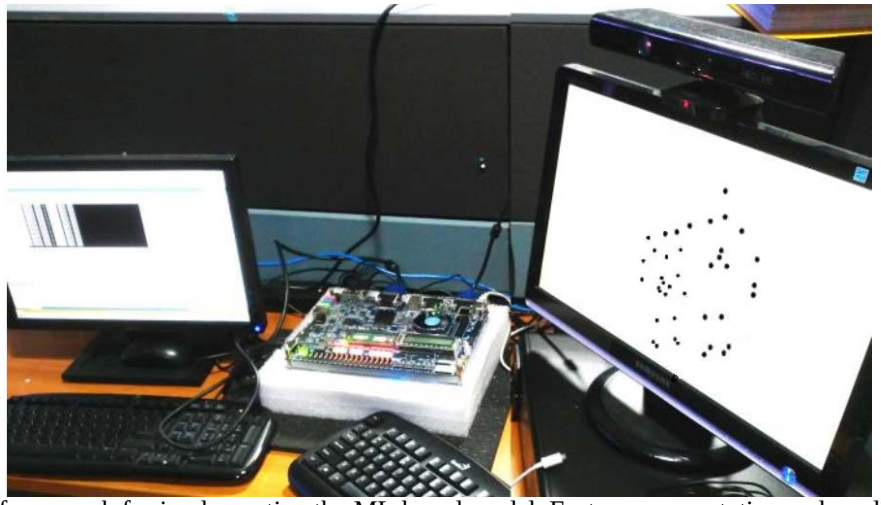

Fig. 8. A framework for implementing the ML-based model. Features computation and prediction are planned to be designed in an FPGA device

References

[1] Janoch A, Karayev S, Jia Y, Barron JT, Fritz M, Saenko K, Darrell T, "A category-level 3d object dataset: Putting the Kinect to work," *Consumer Depth Cameras for Computer Vision, Springer*, pp 41–165, 2013. Article (CrossRef Link)

[2] Girshick R, Shotton J, Kohli P, Criminisi A, Fitzgibbon, "An Efficient regression of general-activity human poses from depth images," in *Proc. Of 2011 IEEE International Conference on Computer Vision*, pp 415–422, 2011. Article (CrossRef Link)

[3] Shotton J, Sharp T, Kipman A, Fitzgibbon A, Finocchio M, Blake A, Cook M, Moore R, "Real-time human pose recognition in parts from single depth images," *Communications of the ACM 56*, PP 116–124, 2013. Article (CrossRef Link)

[4] Sun M, Kohli P, Shotton J, "Conditional regression forests for human pose estimation," in *Proc. of IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, IEEE*, pp 3394–3401, 2012. Article (CrossRef Link)

[5] Haque A, Peng B, Luo Z, Alahi A, Yeung S, Fei-Fei L, "Viewpoint invariant 3d human pose estimation with recurrent error feedback," *arXiv preprint*, 2016. Article (CrossRef Link)

[6] Ferguson D, Silver D, "Pose estimation using long range features," US Patent 9255805, 2016. Article (CrossRef Link)

[7] Belagiannis V, Zisserman, "A Recurrent human pose estimation," *arXiv preprint*, 2016. Article (CrossRef Link)

[8] Carreira J, Agrawal P, Fragkiadaki K, Malik J, "Human pose estimation with iterative error feedback," *arXiv preprint*, 2015. Article (CrossRef Link)

[9] Gavrila D, Davis L, "Towards 3-d model-based tracking and recognition of human movement: a multi-view approach," *In: International workshop on automatic face-and gesture-recognition*, pp 272–277, 1995. Article (CrossRef Link)

[10] Campbell LW, Bobick AE, "Recognition of human body motion using phase space constraints," in *Proc. of Proceedings IEEE Fifth International Conference on Computer Vision*, pp 624–630, 1995. Article (CrossRef Link)

[11] Yacoob Y, Black MJ, "Parameterized modeling and recognition of activities," in *Proc. of Sixth IEEE International Conference on Computer Vision*, pp 120–127, 1998. Article (CrossRef Link)

[12] Husz ZL, Wallace AM, Green PR, "Behavioural analysis with movement cluster model for concurrent actions," *Journal on Image and Video Processing*, 2011. Article (CrossRef Link)

[13] Ali S, Basharat A, Shah M, "Chaotic invariants for human action recognition," in *Proc. of 11th IEEE International Conference on Computer Vision ICCV 2007*, pp 1–8, 2007. Article (CrossRef Link)
[14] Tran KN, Kakadiaris IA, Shah, "SK Modeling motion of body parts for action recognition," *In: Citeseer BMVC*, vol 11, pp 1–12, 2011. Article (CrossRef Link)
[15] Yun K, Honorio J, Chattopadhyay D, Berg TL, Samaras D, "Two-person interaction detection using body-pose features and multiple instance learning," in *Proc. of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp 28–35, 2012. cArticle (CrossRef Link)
[16] Hu T, Zhu X, Guo W, Su K, "Efficient interaction recognition through positive action representation," *Mathematical Problems in Engineering*, 2013. Article (CrossRef Link)
[17] Wolf C, Lombardi E, Mille J, Celiktutan O, Jiu M, Dogan E, Eren G, Baccouche M, et al., "Evaluation of video activity localizations integrating quality and quantity measurements," *Computer Vision and Image Understanding, 127*, pp 14–30, 2014. Article (CrossRef Link)
[18] Bloom V, Argyriou V, Makris D, "G3di: A gaming interaction dataset with a real time detection and evaluation framework," in *Proc. of Workshop at the European Conference on Computer Vision*, Springer, pp 698–712, 2014. Article (CrossRef Link)
[19] Wang K, Wang X, Lin L, Wang M, Zuo W, "3d human activity recognition with reconfigurable convolutional neural networks," in *Proc. of the 22nd ACM international conference on Multimedia*, pp 97–106, 2014. Article (CrossRef Link)
[20] Xu N, Liu A, Nie W, Wong Y, Li F, Su Y, "Multi-modal & multi-view & interactive benchmark dataset for human action recognition," in *Proc. of the 23rd ACM international conference on Multimedia*, pp 1195–1198, 2015. Article (CrossRef Link)
[21] Alhammami M, Ooi CP, Tan WH, "Violent actions against children," *Data in Brief*, 12, 480 – 484, 2017. Article (CrossRef Link)
[22] Liang B, Zheng L, "A survey on human action recognition using depth sensors," in *Proc. of 2015 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pp 1–8, 2015. Article (CrossRef Link)
[23] Yang X, Tian YL, "Eigenjoints-based action recognition using naive-bayes-nearest-neighbor," in *Proc. of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, pp 14–19, 2012. Article (CrossRef Link)
[24] Wang J, Liu Z, Wu Y, Yuan J, "Mining actionlet ensemble for action recognition with depth cameras," in *Proc. of 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012. Article (CrossRef Link)
[25] Luo J, Wang W, Qi H, "Group sparsity and geometry constrained dictionary learning for action recognition from depth maps," in *Proc. of the IEEE International Conference on Computer Vision*, pp 1809–1816, 2013. Article (CrossRef Link)
[26] Vemulapalli R, Arrate F, "Chellappa R, Human action recognition by representing 3d skeletons as points in a lie group," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 588–595, 2014. Article (CrossRef Link)
[27] Xia L, Chen CC, Aggarwal JK, "View invariant human action recognition using histograms of 3d joints," in *Proc. of 2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, pp 20–27, 2012. Article (CrossRef Link)
[28] Zhao X, Li X, Pang C, Zhu X, Sheng QZ, "Online human gesture recognition from motion data streams," in *Proc. of the 21st ACM international conference on Multimedia*, ACM, pp 23–32, 2013. Article (CrossRef Link)
[29] Zanfir M, Leordeanu M, Sminchisescu C, "The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection," in *Proc. of IEEE International Conference on Computer Vision*, pp 2752–2759, 2013. Article (CrossRef Link)
[30] Wu D, Shao L, "Leveraging hierarchical parametric networks for skeletal joints based action segmentation and recognition," in *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*, pp 724–731, 2014. Article (CrossRef Link)

[31] Muller M, R ¨ oder T, Clausen M, "Efficient content-based retrieval of motion capture data," *ACM Transactions on Graphics (TOG),* ACM, vol 24, pp 677–685, 2005. Article (CrossRef Link)
[32] Alhammami M, Ooi CP, Tan WH, "Violence recognition using harmonic mean of distances and relational velocity with k-nearest neighbour classifier," *International Visual Informatics Conference,* Springer, pp 132–139, 2015. Article (CrossRef Link)
[33] Hall MA, "Correlation-based feature selection for machine learning," Ph.D. thesis, The University of Waikato, 1999. Article (CrossRef Link)

Dr. Muhammad is a Ph.D. graduate in the Faculty of Engineering (FoE), Multimedia University, Cyberjaya, Malaysia. He received the MSc. Degree in Integrated Systems and Circuits Design (ISCD) from Carinthia University, Villach, Austria, in 2010. His research interests include Integrated Sensors and Systems, Digital Design, FPGA and SoC, Embedded Systems, Machine Vision, Machine and Deep Learning, and Information and Communication Technology. He has published two regular papers in Data in Brief and Journal of Engineering and Applied Sciences, and two conference papers in the 4th International Visual Informatics Conference and 2015 IEEE Conference on Sustainable Utilization and Development in Engineering and Technology (CSUDET).

Dr. Samir has a Ph.D. in management information systems (2010) and a M.Sc. in computer information systems (2005) and his background is bachelor in electronic engineering (1998) from Damascus University. He is an Assistant Professor at Dhofar University (DU) located in Oman. He has more than eighteen (18) years of industrial experience, research, and teaching. He teaches several courses for PG, BA and UG levels at Dhofar University. Samir has received the best Faculty member award in the College of Commerce and Business Administration of the AY 2016-2017 at Dhofar University. He is currently serving as a n acting director of the community service and continuing education centre since 2017, chairperson of the centre for entrepreneurship since initiating since in 2014 in addition to heading the Industry Engagement Committee since 2015 to present. Also, he served as an acting chairperson of the management information systems department (summer 14-15 and summer 15-16). Furthermore, he is a mentor in the Omani national mentorship program of the public authority of SME development in Oman 2017/ & 2018.

Dr. Ooi Chee-Pun received his M.Sc. in Electronics from Queen's University of Belfast, UK and a Ph.D. in Engineering from the University of Malaya. He is currently the Chairman of MMU Digital Home & Lifestyle Centre at Multimedia University. Dr. Ooi's areas of expertise include FPGA System Design, Embedded System Design, and IoT application. He has been involved in various government-funded projects since he started his career with the University. His works have been published in numerous international journals and conferences.

Dr. Tan Wooi-Haw received his M.Sc. in Electronics from Queen's University of Belfast, UK and a Ph.D. in Engineering from Multimedia University. He is currently a senior lecturer at Multimedia University. Dr. Tan's areas of expertise include image processing, embedded system design, and computer networking. He has been involved in various government-funded projects since he started his career with the University. His works have been published in numerous international journals and conferences. Besides, he has also co-authored two textbooks on microcontroller system.