International Journal of Computer Information Systems and Industrial Management Applications.
ISSN 2150-7988 Volume 13(2021) pp. 031-042

Received: 10 May, 2020; Accepted: 26 Feb., 2021; Published: 25 Mar., 2021

Low-cost Design of Vision-based Natural User Interface via Dynamic Hand Gestures

Richa Golash[1], Yogendra Kumar Jain[2]

[1] Samrat Ashok Technological Institute,
Civil Lines, Vidisha, Madhya Pradesh, India 464001
golash.richa@gmail.com

[2] Samrat Ashok Technological Institute,
Civil Lines, Vidisha, Madhya Pradesh, India 464001
ykjain_p@yahoo.co.in

***Abstract*: Advancement in computer vision and pattern recognition fields has opened many new dimensions for object recognition and visual tracking in videos. But vision-based interaction of a human with a machine using moving hand gestures is still in a primitive stage because seamless localization and conversion of irregular trajectory to command in RGB images are challenging tasks. The hand is a non-rigid object with diverse posture shapes, some posture occupies a larger area and some very small. The surface of a hand is uneven thus edges are not very clear during movement. Therefore, algorithms example background subtraction, segmentation using contour information, and skin color detection are not practically applicable in unconstrained background conditions. Hence researchers prefer advanced cameras that can detect the skeleton structure and provide more information apart from RGB values for each pixel of the image frame to facilitate the hand detection stage in dynamic hand gesture recognition (DHGR). The approach of using advanced cameras in DHGR increases the overall cost of interfaces, subjects require technical knowledge to operate them, and hence applications of DHGR are limited to those areas where complexity, cost, and expertise are not very critical factors. The goal of this paper is to propose the low-cost, simple design of vision-based human-machine interaction via dynamic hand gestures that directly work on RGB images, can be easily integrated with any day-to-day machine and the technique is invariant to the user's age and skills. The architecture of the proposed method utilizes Faster Region-based Convolutional Neural Network (Faster R-CNN) for hand spotting and recognition, tracking is accomplished using SIFT (scale-invariant feature transform), modified Backpropagation Artificial Neural Network is final added to efficiently translate hand movement into machine command with optimal computations. The proposed technique is simple yet robust in hand detection and capable to track and interpret a non-rigid object directly in the colored videos, captured in a real-time environment.**

***Keywords*:** Natural-user-interface, Machine learning, Deep Neural network, Faster R-CNN dynamic hand recognition, visual object tracking.

I. Introduction

Hand gestures are the most successful nonverbal mode of communication among human beings and due to this fact, the revolutionary idea proposed by Bolts R. A. 'Put-that-there' is greatly appreciated and still being researched for establishing a natural method for human-machine interaction. With the rapid development in the areas, for example, Computer Vision, Pattern Recognition, Machine Learning, and Image Processing, vision-based interaction of human with machines have been possible, and sign language recognition is one of the benchmark application of hand gesture recognition [HGR] [1-3].

Till now many natural user interfaces (NUIs) have been created using sign languages of hand. The user shows some static postures of hand, machine capture the image through an embedded camera and interprets that image as one of command according to the training given. Bergh M. et al. [4], used the pointing direction of hand gestures to control robot movement. Similarly, Ren Z. et al. [5] developed a vision-based arithmetic computation tool system and Rock-Paper Scissor game using hand postures. Ohn-Bar E. et al. [6] designed a vision-based gestural interface to control a car infotainment system. Esfahani E.N. et al. [7] proposed contactless handling of electronic equipment present in the surgical room of a hospital by the doctors to reduce the time of surgery and spreading of contagious diseases. Ravindu H.M. [8] proposed an intelligent wheelchair system to reduce the human assistance required by physically challenged people. The common fact among all this research is that all these techniques mainly involve advanced cameras (e.g., Kinect, Leap Motion) that has inbuilt software to detect hand by applying threshold depth distance in the image. Most of the works are either confined to recognize and classify static posture of a hand or limited to track finger movement of gesture with length 20-40 frames [9-12].

Kinect and Leap Motion are motion-sensing cameras working prominently in the field of HGR. Kinect cameras possess an infrared laser projector with a monochrome CMOS sensor and provide the skeleton of a full body. Leap Motion cameras possess two cameras, three infrared LEDs, and a specially designed wide-angle camera sensor to minutely focus on hand physical structure [9-12]. The performances of

both the cameras are better than the simple RGB cameras because depth data is invariant to light change, thus it becomes easier for researchers to get the fingertip location and other details of a hand using the inbuilt facility of these cameras. But the easy process of hand detection comes along with the cost, complexity [12], and the requirement of skilled users. Though with the advent of micro technology the cost of cameras has been reduced drastically, in comparison with simple RGB cameras, advanced cameras are still 15 times costlier. Moreover, maintenance, regular calibration as per user hand shape and posture are some of the unacceptable factors in designing general-purpose vision-based human-machine interfaces. Hence in terms of affordability and usability design of Natural Human-Machine Interface (NHMI's) with simple RGB cameras are more likely a good option. This paper aims to propose a low-cost design of a vision-based natural user interface via dynamic hand gestures. Since the methodology utilizes a webcam which is a very familiar camera among all age group users, therefore this vision-based user interface design is easy to understand, trouble-free to integrate with any day-to-day machine.

Natural User Interfaces via dynamic hand gestures are real-time applications of visual object tracking. Unpredicted background, illumination variation, uneven surface of hand region, speed variation during movement, and above all 2D image representation of 3D motion, all these physical and behavioral characteristics of hand is accompanied by the loss of information when the dynamic hand gestures are captured using simple camera [13]. Mainly there are two categories of approaches in techniques that are directly working on RGB images. The first category belongs to those researchers whose works deal with static hand postures and in the second category, we have discussed those techniques where any type of hand motion is taken.

Chen Z. H. et al. presented hand gesture recognition on a single black background, the technique comprises of background subtraction algorithm followed by finger and palm segmentation, the obtained hand shape is then classified using a rule-based classifier [14]. Similarly, Simion G. et al. [15], performed background subtraction to detect hand region, since information is lost during this process, they suggested maintaining a codebook for the background model. Further, they used the multi-scale version of the Mean-shift algorithm (MSMF) to detect the position of fingers and palm region. In this technique, experiments are conducted on a limited number of postures [15].

Pei Xu et al. [16] proposed vision-based mouse-cursor control, according to them, a robust system should not respond to the transient stages arising in between the change from one posture to another. To respond only stable points author suggested many preprocessing steps for example background segmentation, filtering of skin color, removing blur, contour extraction, centering the hand region, etc. before feeding the binary image to CNN modified LeNet-5 [16]. Similarly, Pinto F. [17] used binary images to train the CNN network. Here binary images are obtained after color segmentation using a small neural network architecture with two hidden layers [17].

Singha et al. [18] technique has shown good results in tracking dynamic hand gestures in real-time background. In this method after removing the face by applying the Viola-Jones algorithm, they have performed skin segmentation on the gray and colored difference images. Features are extracted and tracked using Kanade–Lucas–Tomasi Tracker, but by virtue, KLT feature decreases subsequently, thus authors have added compact criteria, CAMShift algorithm, and regenerated the features after every 30 frames. Tracking is performed based on 44 features extracted during the process [18].

For object tracking in RGB images, local features show high precision due to their deep understanding of spatial information of the pixel. He Wei et al. [19], proposed a generative model between the motion of local features and the global motion of an object. Bao J. et al. [20], too agreed that the accuracy of hand gesture segmentation is largely affected by skin color and a high degree of freedom in hand geometry, therefore local features, SURF show more capability in recognizing an object in the unconstrained background. But in [20], it is assumed that the hand occupies a large area of the frame, this work was extended by Yao Yi et al. [21] by using adaptive SURF tracking. According to [21], the exact location of the gesturing hand is not required always. To remove redundant features and calculation of displacement of hand region is done, using pruning process based on threshold displacement. But the threshold displacement range was limited to 3 to 40 pixels, which increases the ROI region, and the technique was applied to binary images.

In the design of natural interface via dynamic hand gestures, semantic development between trajectory and command is a very essential component, which is generally overlooked by the researchers. Since the same hand gesture does not follow the same path, this factor is also one of the main reasons that no rules can be developed for grammar formation between trajectory and instruction given to a machine. Dinh D. et al. [22] developed an interface for smart home appliances using four depth silhouettes of hand and trained them using the Random Forest classifier. Not much tracking has been done in their techniques. Kılıboz N. et al. [23] proposed trajectory-based human-computer interaction, a total of 11gesture based commands are created by moving hand in XY direction. Their technique resembles more glove-based techniques because motion is captured using a magnetic motion tracking device attached to the user's hand. Tran D. et al. [24] proposed human-machine interaction via fingertips tracking of seven hand contours extracted using Kinect V2. The sampling gesture was confined to 20-45 frames collected at a speed of 30 frames per second. Zeng J. et al. [25] developed NUI based on hand gestures for special people. They used a multi-cue system with frame-based motion history images, but the motion was limited to the abduction and adduction process of thumb and three fingers in a black fixed background. Grif H.S. et al. [26] used hand pads and color strips to detect hand fingers to control mouse cursor through hand gestures.

The survey conducted on various approaches highlights some major issues that restrict the applicability of low-cost RGB cameras in the design of vision-based user interface via dynamic hand gestures are:

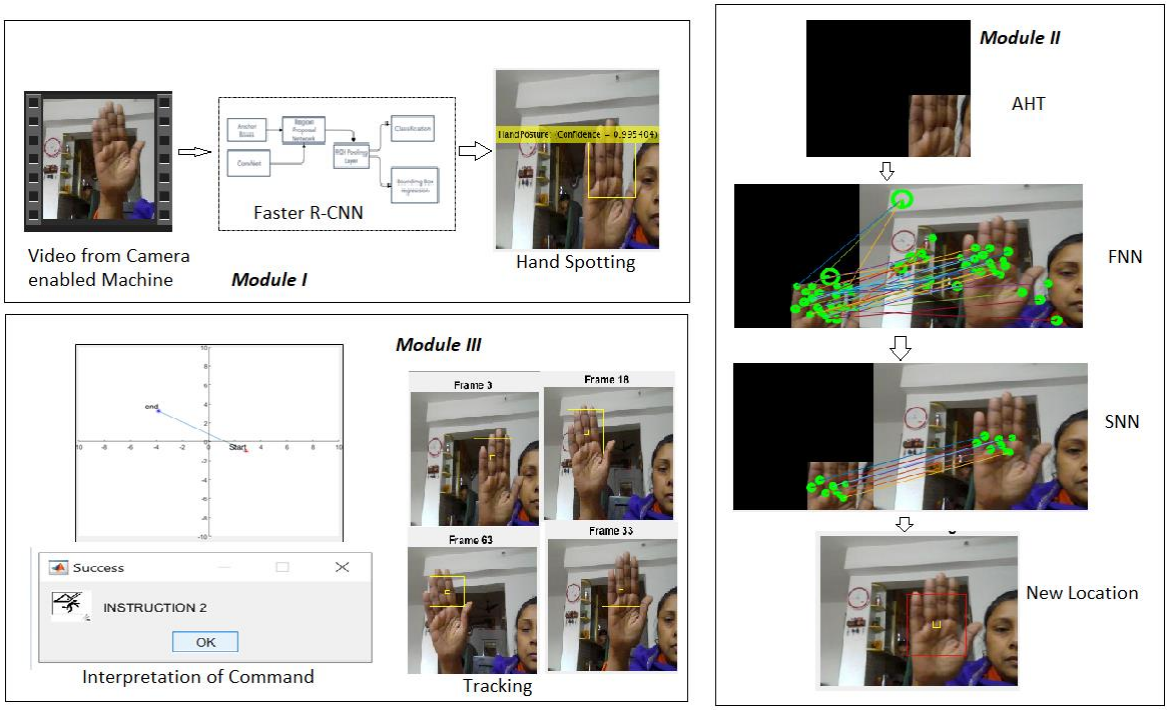


Figure 1. The architecture of the proposed system

(i) Hand is a non-rigid biological object, the surface of the hand is irregular, its motion is complex, thus loss of information due to self-occlusion, scale change, blur, change in camera-view is very prominent in RGB images of hand movement.

(ii) Hand has a very high degree of freedom in contrast with other commonly tracked objects. Because of this, even the same hand posture is not identical in shape, additionally, edges are unclear in the images when a hand is moving. Thus, hand segmentation, contour extraction, skin color detection, could not give good results in real-time application i.e., NUI.

(iii) User wants complete freedom in terms of posture and background selection. Thus, in the design of NUI's background subtraction or conversion to binary images is not possible always.

This paper aims to develop a robust system that is capable to handle hand detection as well as tracking in a real-time background, and finally interpreting the trajectory as a machine command. In the proposed method hand is spotted and detected using Faster RCNN, the template obtained in this process act as an active hand template of the moving hand in that video sequence. After every 10th frame AHT is updated. Next, we detect the Scale Invariant Feature Transform of AHT and called them Active-SIFT features, and use these features to recognize the new location of hand in subsequent frames. Instead of taking SIFT features of a complete initialization frame, we select features of only the active region for matching and recognizing the hand region in subsequent frames. The advantage of using this strategy is since the area of the moving hand region is very small and consequently, the numbers of SIFT features are very less therefore the time and computations in matching and pruning unnecessary features are saved. The coordinates obtained in every frame are plotted as a trajectory using the Cartesian plane system. The cartesian points of the trajectory are matched using an Artificial Neural Network (ANN) trained by a modified-Backpropagation algorithm to interpret machine command. In this work, we have designed eight trajectory-based general-purpose machine commands (INSTRUCTION 1-8) for vision-based machine interaction. The distinguishing features of the proposed technique are that it enables us to derive the trajectory without performing background subtraction or applying any type of segmentation process. It applies to four postures and can be extended to many. The command interpretation step is invariant to the trajectory of movement.

II. Architecture of Proposed system and Algorithm

Figure 1 illustrates the systematic working of the proposed methodology. It is divided into three modules as follows:

Module I, deals with the acquiring of video of length 100-150 frames, preprocessing of raw data. In this module, we determine the hand posture used by the user, which we call an active hand template (AHT) and point of start (POS) from where the user has started its hand gesture. The dimensions of

AHT are calculated and a bounding box of that value is created at POS.

Module II handles the tracking part of the design. Here the local features of the hand template (HT) which are very small in number are used to search the new position of the moving hand region in successive frames. The location of the dominant movement vector is determined and updated as the centroid of movement.

Module III is the execution of the command module. Here, the centroid of all frames is accumulated and plotted using the quadrant system of the Cartesian plane. This plot gives the trajectory of hand gestures in the test data sequence known as the test trajectory. The test trajectory is matched with the trajectory of the trained database and converted to the command using modified backpropagation Artificial Neural Network. Next, we discuss the mathematical understanding of the algorithm.

A. Module I: Hand Modelling and Spotting

In the design of NUI via dynamic hand gesture Hand Modelling and Spotting plays a very significant role because the accuracy of this stage decides the efficiency, suitability, and applicability of NUI in a real-time scenario. Hand modeling means finding the hand posture that is to be tracked and spotting means, the frame number from which the actual hand movement starts. Three major issues that are mostly encountered while tracking a hand motion are

(i) A hand is a versatile object in comparison with other objects and the area occupied in the image frame is dependent on the posture selected.

(ii) It is not fixed that the subject starts the motion from the first frame or the fixed position in the frame.

(iii) Scale variation is very prominent in videos of hand movement. The model selected must cope with the variation.

In the proposed method the above-mentioned issues are solved using Faster-Region-based Convolutional Neural Network (Faster-RCNN) a deep neural network architecture (DNN) is a modified version of region-based convolutional neural network (R-CNN). Faster R-CNN is a generic object detection scheme, that can provide a fast and accurate position of the object, represented in a bounding box with a confidence score [27]. The core architecture consists of two networks a Region Proposal Network (RPN) and a Fast R-CNN module [28], combined into one network by merging their convolutional features. The important blocks (as shown in figure 2) of Faster R-CNN are discussed as follows:

Region Proposal Network (RPN): RPN is an independent ConvNet designed to generate region proposals directly instead of using an edge box algorithm, by changing scales and aspect ratio known as anchor boxes. Anchor boxes are bounding boxes with predefined height and width to capture the scale and aspect ratio of the target object. For every tiled anchor boxes, the RPN predicts the probability of object, background, intersection over union (IoU) values. The advantage of using the anchor box approach over sliding window-based detector in object detection is that former detect, feature encode and classify the object in a single process [27].

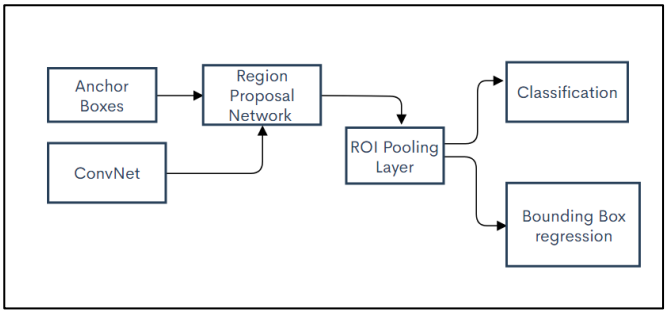


Figure 2. The architecture of Faster R-CNN

ROI Pooling Layer: The region proposal generated by the RPN is of variable size. To classify objects we require fixed-size inputs therefore outputs from RPN are passed into the ROI pooling layer, which uses max pooling to convert features of valid anchors or into fixed size (HxW) feature map. In this layer, the number of outputs is equal to the number of inputs. Each valid ROI is specified by its index number, coordinates of its top-left corner (x, y), height h, and width w .

Bounding box regression layer: In this section of the network a new fully-connected (FC) layers with two branches replace the older FC. One branch is responsible for bounding box predictions. The second branch is the softmax classifier and is responsible for generating its class and score of objectness. This layer refines the locations of the bounding box by using smooth L1 loss.

Loss Function: The faster R-CNN [27] is performing two task one is classification, and the second is bounding box regression thus, multitask loss in faster R-CNN is given as equation 1:

$$Total\ Loss = Class.loss + BB\ reg.loss \quad (1)$$

Class. Loss is a log classification loss calculated over, object vs no object. It P_j is the predicted probability of jth anchor to be an object computed by the softmax layer and P_j^* is the ground truth of anchor j belongs to the object or not. Then classification loss is given by (equation 2):

$$Class.Loss(P_j, P_j^*) = -P_j^* log P_j - (1 - P_j^*) \log (1 - P_j) \quad (2)$$

Here $P_j^* = 1$ if the anchor is an object and $P_j^* = 0$ if an anchor is not an object. BB reg. loss is bounding-box regression loss and it is calculated as the loss over true BB regression targets for a particular class on the predicted BB regression. If T_j^* is the tuple of ground truth box associated with positive anchor and T_j is the tuple of predicted bounding box, then BB regression loss is given as (equation 3):

$$BBreg.Loss(T_j, T_j^*) = Y(T_j - T_j^*) \quad (3)$$

Here Y is the robust smooth L_1 loss function.

B. Module II: Motion Modeling

Motion Modeling means finding the trajectory of hand motion from start till the end. In the proposed method to find the target hand posture (AHT) in the subsequent frame, we have used the Scale-Invariant Feature Transform (SIFT) Algorithm designed and described by David Lowe in our tracking process [29]. SIFT features have a high distinctiveness and better detection accuracy toward local image distortions, viewpoint change, and partial occlusion and are helpful in real-time fast-tracking of the target [29], [30].

SIFT algorithm comprises of feature detector as well as the feature descriptor. In general, features are high contrast areas example edges, in the image. These features are extracted such that they are detectable even in noise, scale variation, and changes in illumination. Each feature is located by four parameters: $f_i = \{p_i, \sigma_i, \varphi_i, gh_i, d\}$, where $p_i = (x_i, y_i)$ is the 2D position of SIFT keypoint, σ_i is the scale, φ_i gradient orientation within the region and d is a 128-dimensional descriptor of the key point i.

In our approach, we find the SIFT features of the AHT template obtained in module I. Since it contains only the target hand posture and is small as compared to the image frame [240, 240]. Therefore, this approach saves the time of matching unnecessary features and pruning them further [20], [21].

In this process target object recognition and localization in the subsequent frames of a video is accomplished in two stages: In this first stage, we determine the first nearest neighbors (FNN) of all the features of the AHT template in the current frame using minimum sum squared difference method. An experiment (as shown in figure 3) is carried out in an indoor environment, here a subject is performing her hand gesture movement inside a room, the AHT of this video as shown in figure 3(a) is determined using module I. Let there are *m* key features in AHT frame, given as $S_{AHT} = \{f_i\}^m$, where f_i is the feature vector at i^{th} location. Let $S_{cur} = \{f_j\}^k$ are k numbers of SIFT features in the current frame, where f_j is the SIFT feature at j^{th} location. We use the best-bin-first search method that identifies the nearest neighbors of AHT features with current frame features. The First Nearest Neighbors (FNN) are defined as the pairs of key points with a minimum sum of squared differences for the given descriptor vector (equation 4).

$$distance\ (a_{AHT}, b_{cur}) = \sqrt{\sum_{i=1}^{128}(a_i - b_i)} \tag{4}$$

where a_{AHT} and b_{cur} are descriptor vectors of features in AHT and current frame respectively. Figure 3(b) AHT frame is concatenated with the current frame and colored lines show the FNN match pairs between them. We observe that there many false and ambiguous or multiple match pairs between them. To improve matching, we perform Lowe's second nearest neighbor (SNN) test (equation 5) and further geometric verification test (equation 6) on the keypoints obtained after SNN. The consistent keypoints after SSN and geometric verification test are shown in figure 3 (c).

Second Nearest Neighbor (SNN). This is done by calculating the ratio between the first nearest neighbor distance (FNND) of HT features with current frame features to the second nearest neighbor distance (SNND) of HT features with current frame features.

$$\frac{distance(a_{AHT},\ b_{cur})}{distance\ (a_{AHT,}\ \ c_{cur})} > 0.8 \tag{5}$$

$$\begin{bmatrix} x^* \\ y^* \end{bmatrix} = vR(\alpha)\begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} T_x \\ T_y \end{bmatrix} \tag{6}$$

Here a_{AHT}, are key points in AHT b_{cur}, $and\ c_{cur}$ are two close neighbors of a_{AHT}. v is isotropic scaling, α is rotation parameter, (T_x, T_y) are translation vector for (x, y). All the consistent key points are shown in figure 3(c). Out of all the close matches in figure () we find a unique key feature that has the least distance among all. That unique key point is selected as a centroid of motion for that frame.

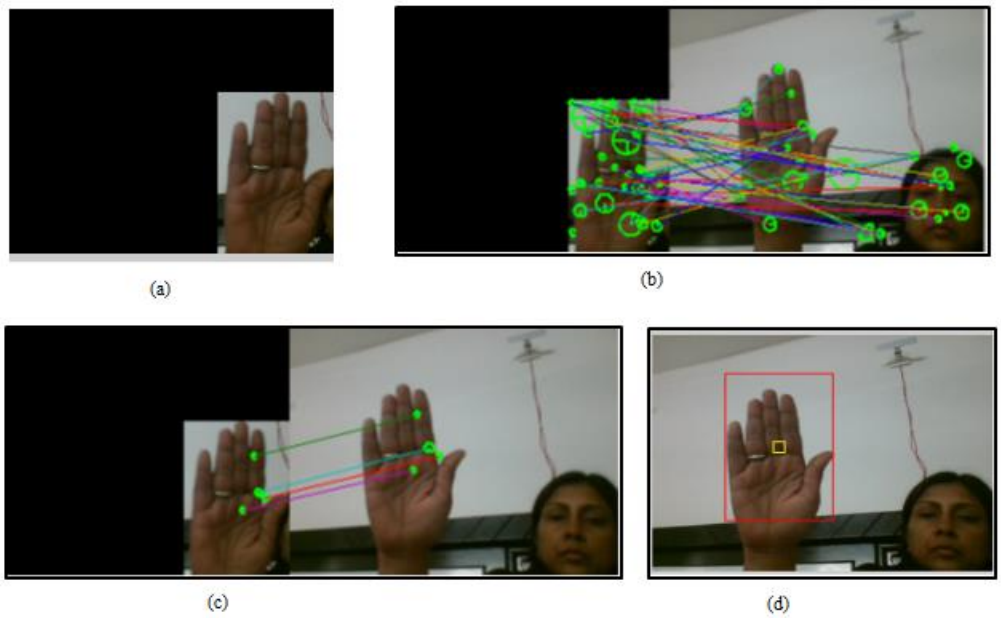


Figure 3. Outcomes of Module II (a) AHT frame (b) FNN match pairs between HT frame and current frame (c) SNN match pairs between AHT frame and current frame (d) Red color box is the new location of the moving hand region and yellow color box is the centroid of movement in the current frame.

C. Module III: Machine learning Algorithm

Module III is the most critical stage in the development of NUI via dynamic hand gestures because the pattern of hand movement is non-linear. The same pattern of hand movement may follow a different path, researchers [23] preferred mechanical devices attached to the hand to record the hand movement. Thus, semantic development between trajectory and machine command is challenging. To resolve the above-mentioned issue, we have divided the image frame into quadrants as a cartesian plane system. We have taken that a subject will start either from the left or the right part of the image frame and can move his hand to either of the four quadrants. In our prototype, we have designed eight visible commands INSTRUCTION 1-8 as shown in figure 6. These commands are developed keeping in mind that any general-purpose machine, especially used in a home environment is operated by 6-10 commands

To classify trajectory to command we have used modified backpropagation of Artificial Neural Network (BP-ANN) as shown in figure 4. The artificial neural network (ANN) is a pragmatic model that can find the pattern buried in data in a quick and precise manner. Backpropagation (BP) is a supervised feed-forward network that reduces the error through the gradient descent rule. In the proposed architecture we have taken four input neurons and eight output neurons and one hidden layer with 10 neurons.

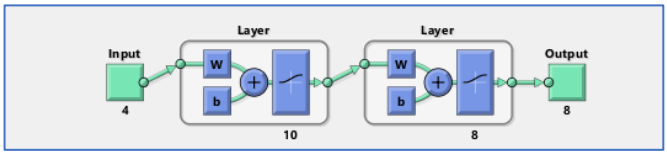


Figure 4. The architecture of the ANN in the proposed method

In traditional BP the performance parameter is highly dependent on learning rate, thus we have used adaptive learning rate η and a momentum factor that increases the previous weight by a factor m ($0 < m < 1$) as given by equation (7). Figure 5 shows the performance of train data the sum of squared error is 0 at 125 epochs [31], [32].

$$\Delta w_{ij}(t) = \eta \delta_j a_i + m \Delta w_{ij}(t-1) \quad (7)$$

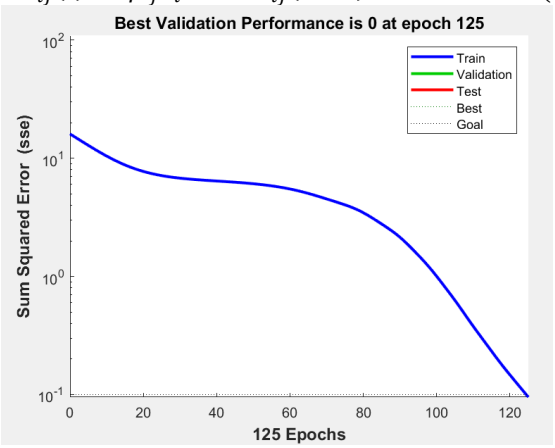


Figure 5. Performance graph of ANN used in the proposed method.

III. Experiments & Results

This section is divided into two subsections, the first subsection evaluates the proposed method, and the second sub-section deals with the comparison of the proposed method with contemporary techniques.

A, Evaluation:

In this section, we will demonstrate the various outcomes for testing and evaluation of our proposed method. We have collected more than 100 videos (as shown in figure 7) from a low-cost camera compatible with Windows WDM having 1280 x 720 pixels image resolution. The dataset comprises videos of hand movement performed by six subjects of 3 age categories: two children (age 10-16 years), two adults (age 20-40 years), and two seniors (age 65 years) with four different hand postures. These videos are capture in simple as well as in complex backgrounds, at different scales and different illumination conditions. The proposed methodology is analyzed on the three parameters discussed by Yang et al. [13]: robustness, adaptivity, and real-time processing.

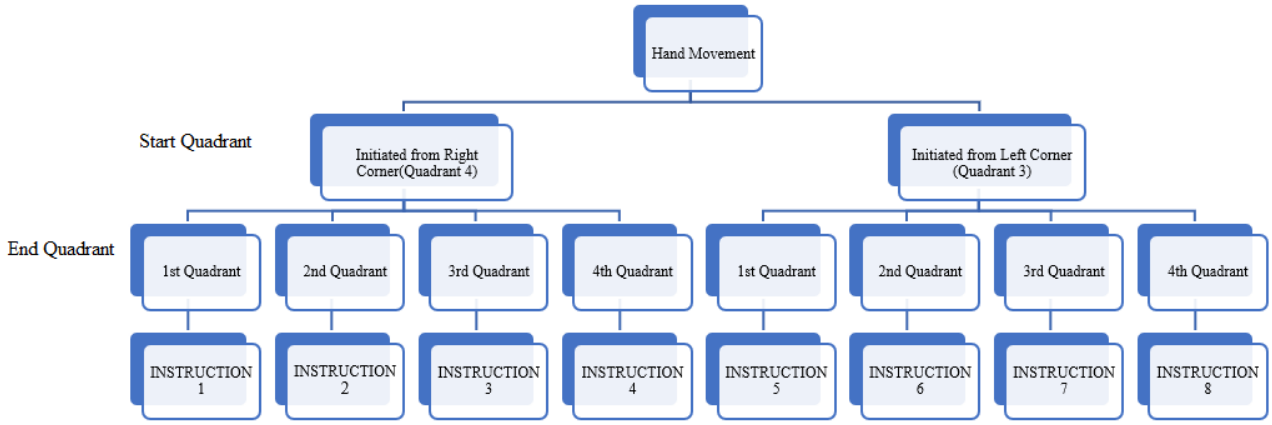


Figure 6. Semantic between hand movement and machine command

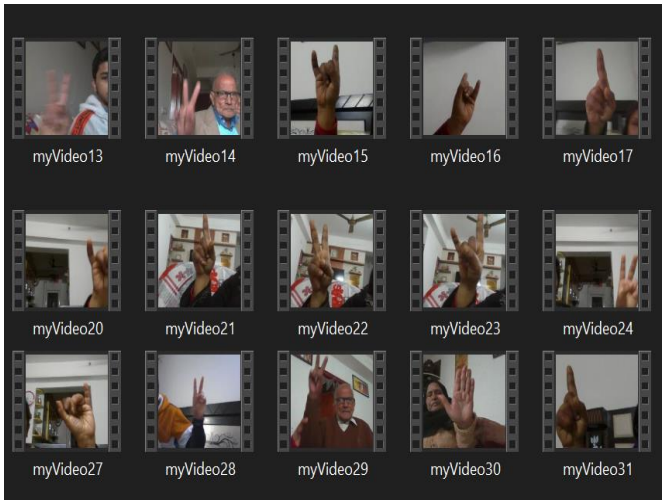


Figure 7. Collection of videos for experiment and evaluation.

1) Robustness:

It is the measure of the system efficiency for tracking simple as well as complicated hand postures, smoothly in the unconstrained background. To test and evaluate we have selected commonly used four hand postures to perform hand movement. The outcomes of the module I are as shown in figure 8 and figure 9.

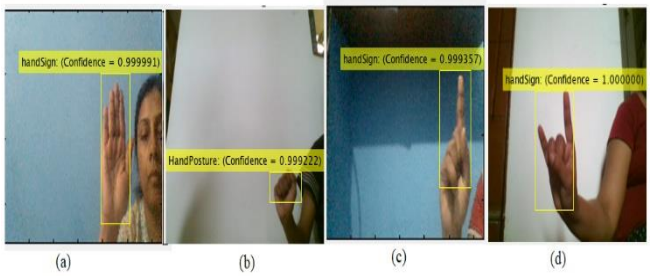


Figure 8. Stage I detection of four different hand postures.

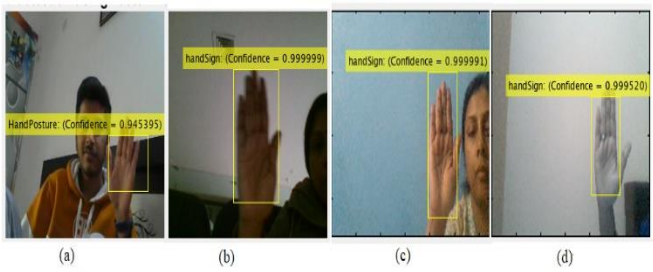


Figure 9. Results of module I in different background and illumination conditions.

The faster R-CNN network of our proposed system is trained on more than 200 images captured in different backgrounds. Figure 8 shows the outcomes of stage I for different hand postures detected by faster R-CNN. For each hand posture, we have collected near 25 videos, where 10 videos are taken in the simple and well-illuminated background, 8 videos each for complex background, 4 videos for cluttered, and 3 videos in less illuminated background conditions. Figure 9 demonstrates the output for hand posture 1 in different

background and illumination conditions. Figure 10 shows the efficiency of the system to detect four different hand postures in a different environment. The results reflect that if the illumination is proper then the detection efficiency of module 1 for all four hand postures is very high but if the illumination is poor then it is difficult to recognize posture 3 and posture 4.

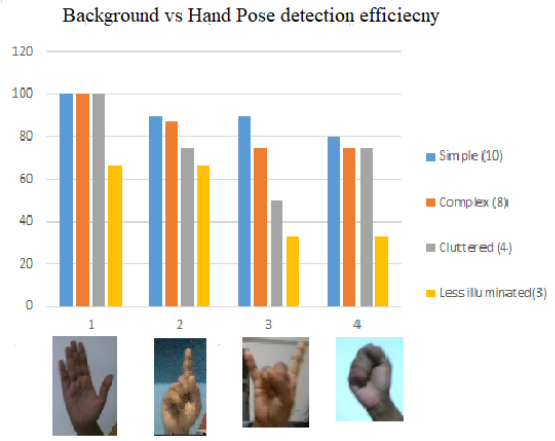
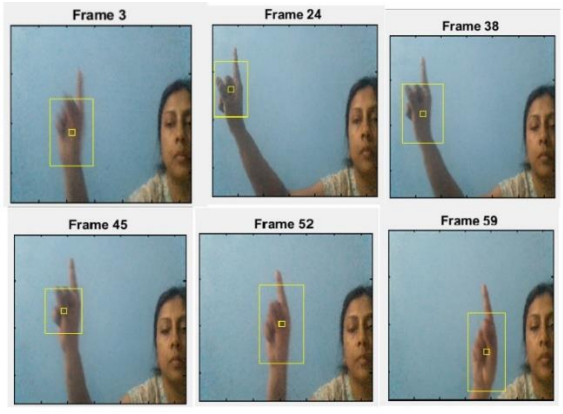


Figure 10. Background vs hand detection efficiency for four-hand postures.

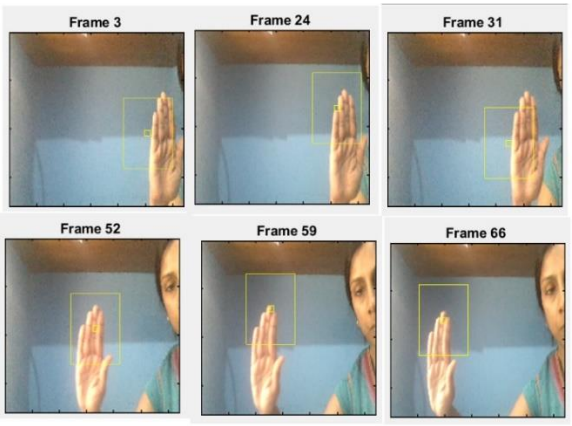
2) Adaptability:

It is a very important factor in the case of hand tracking, as the images of the hand are greatly affected by camera view angle, speed variation, lighting conditions, and scale variation because of the versatile nature of the hand skeleton. Thus, for steady hand detection and tracking, the system should adapt to the aforementioned factors in optimum time and computation. In this research work, we have collected videos of variable length ranging from 80-150 frames to test the performance of the system. Figure 11 demonstrates the tracking results of module II in different environments, example figure 11 (a) & (b) shows the results when the hand view is changed during movement, figure 11 (c) shows the tracking results when the subject could not stabilize its hand in initial few frames (in some frames image of the hand is a blur). The proposed methodology has obtained the tracking accuracy of 98.7% (calculated using equation 8)

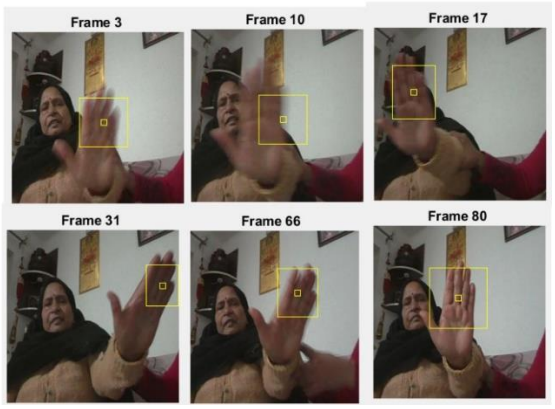
$$Tracking\ Accuracy = \frac{No.of\ correctly\ tracked\ frames}{Total\ No.of\ frames\ in\ a\ video} \times 100 \quad (8)$$



(a)



(b)

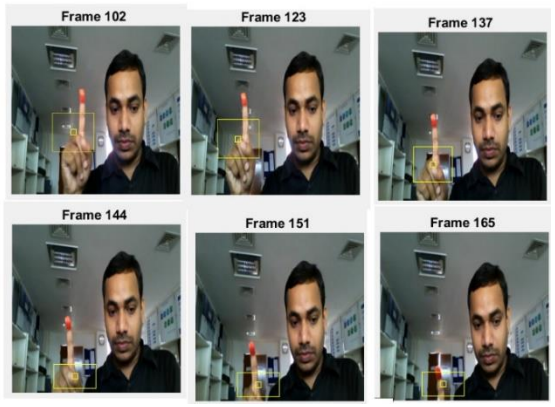


(c)

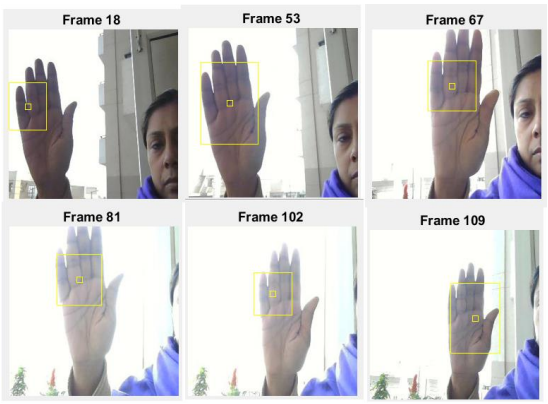
Figure 11. Outcomes of the Hand Tracking stage for different hand postures in different backgrounds.

3) Real-time Processing

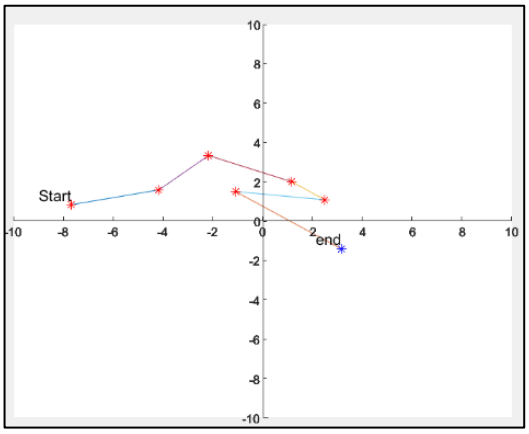
This factor is the building block of NUI, i.e., any system designed must work in maximum possible practical conditions. We have tested our method on publicly available videos of hand movement [18] and some randomly captured videos. the outcomes of our method are illustrated in figure 12, here figure 12(a) & (b) shows the hand detection and localization in the respective frame of that particular video, and figure 12(c) & (d) is the trajectory of hand movement, and its corresponding interpretation as a command (for video in figure 12(b)) respectively. The results justify the real-time suitability of the proposed method.



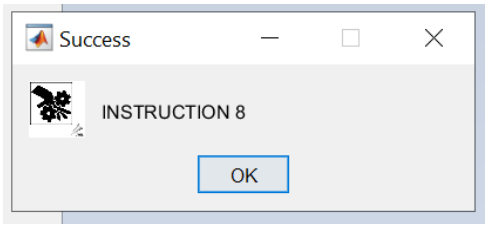
(a)



(b)



(c)



(d)

Figure 12. Results of module II for test data (a) public data sequence [18] (b) real-time data (c) plot of trajectory for real-time data (b), and (d) result of module III.

To test the instruction interpretation efficiency of the system for each hand posture we have collected data for individual commands performed by each hand posture. With the help of equation (9), we find that the accuracy of interpretation of trajectory to instruction. We have captured approximately 20 real-time video sequences of each instruction carried out by four different hand postures in disparate backgrounds. The accuracy of trajectory to command interpretation of hand posture 1 and 2 for all the instructions is 100% followed by hand posture 3 and hand posture 4 with 98 % accuracy. The accuracy of hand posture 3 and 4 are less in poor illumination and highly cluttered background. The average accuracy of the system for interpretation is 99%. Mapping of instruction 5 & instruction 3 with their trajectories are shown in figure 13.

$$Command\ Interpretation\ Accuracy = \frac{\text{No. of Correct Conversions}}{Total\ Sample} \times 100 \quad (9)$$

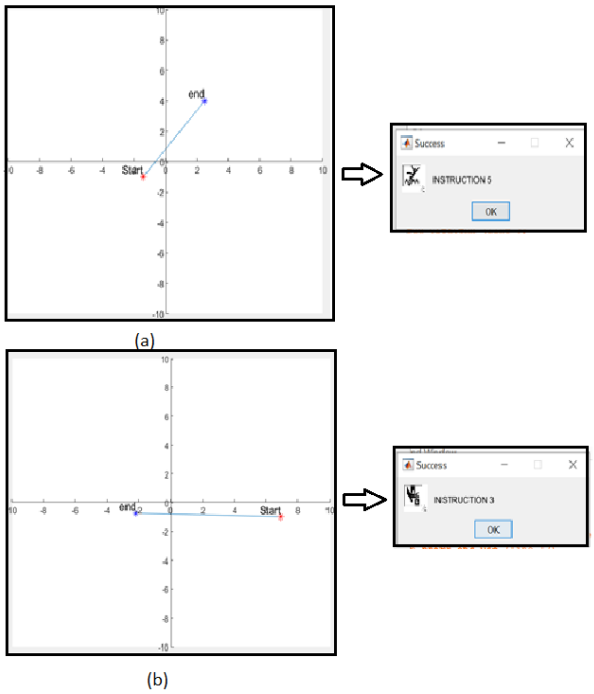


Figure 13. Mapping of trajectory with commands for Instruction 5, **and** Instruction 3 are illustrated. Red * is for the start position and blue * is the end position of the gesture.

B. Comparison with the contemporary techniques:

In this section, we have compared the features of existing techniques that have worked purely on RGB images with the proposed methods [14], [15], [18], [25], [26], [33]. Also, we have discussed some of the research works that describe their technique as 'low-cost' concerning RGB cameras. Table 1 illustrates the hand detection, tracking, and classification strategy of individual techniques. It is observed that the maximum techniques which have utilized RGB images limit their work to static hand gesture recognition [14], [15], [26]. The support system developed by Zeng, J et al. [25] for people with brachial plexus injuries, the motion is limited to the abduction and adduction process of thumb and three fingers

in a black fixed background. Premaratne, P et al. [33] and Singha, J. et al. [18] have proposed tracking in RGB images using the Lucas-Kanade tracking algorithm. In KLT tracking feature points decreases in subsequent frames and the calculation involved in matching requires a longer time. To overcome these issues [18] acquired a frame in every 100 ms. and [33] carried out a recognition process after every 30th frame. To increase the efficiency of the system [18] added the CAMshift algorithm, minimum eigenvalue, and compact criteria. Increasing the number of algorithms though makes the tracking algorithm but makes the methodology complex and practically unrealizable. In contrast to the techniques discussed our proposed method is simple and robust, it is purely based on dynamic hand gesture recognition. The methodology puts no constrain on the background or the length of the gesture. The different results confirm that detection using Faster R-CNN and tracking using the SIFT algorithm gives better efficiency with less complexity in true color images. The trajectory to command interpretation method is also a unique feature of this technique, which has enabled to design of a low-cost vision-based interaction of humans with a machine.

Vandayar N. et al. [34] also suggested a low-cost system of HGR, the main limitation of the design is that it considers only a single hand pose and applicable only for one user. Bautista A.G. C. et al. [35] put forward a case study of a low-cost and fast hardware system on hand gesture recognition, but in their design, they used three cameras and the background was limited to black to perform a hand-arm movement. In this technique for finding wrist and finger position Y variation and Convex hull method are used. According to the researchers still, a confusion state exists between wrist and finger segmentation when the hand is not fully visible. Thabet E. et al. [36] proposed a low-cost skin segmentation technique for dynamic hand gesture recognition. They highlighted that YCbCr can be a good choice as the process of converting RGB to YCbCr is simple, also by putting threshold range on Cb and Cr component skin color objects can be detected easily. But this paper does not discuss the presence of other skin-colored objects in real-time conditions.

Research Paper	Hand Detection tech.	Tracking Features	Classification Strategy	Remarks
Chen Z et al. [14]	Background subtraction	Worked on static images	Rule Classifier	Complex calculations to find palm centre point. Motion Parameter not used. Other skin color objects degrade the performance
Simion G. et al. [15]	Background subtraction	Worked on static images	Mean-shift with multimode algorithm	Six machine commands are generated using static combination of palm and finger. Many parameters have fixed values.
Grif, H.S. et al. [26]	Hand is detected using color strips	No tracking	One to one mapping	Only three mouse events are produce
Zeng, J et al. [25]	Black fixed background	Motion history image	State transition network	Motion is limited to the abduction and adduction process of thumb and three fingers.
Premaratne, P et al. [33]	Skin segmentation	Lucas-Kanade	Neural Network	Image is captured in every 100ms. Technique is carried out on binary images. Gesture length is ver small.
Singha, J. et al. [18]	Frame difference and skin filtering	KLT feature+CAMSHIFT	Neural Network	Total 44 features are selected for tracking. Complex algorithm for tracking.
Proposed Method	Frame difference, connected component	AcSIFT features	Neural Network	Detection and tracking in real-time environment. Low-cost with easy understandable esign to integrate with any day-to-day machine.

Table 1: Comparative study of the proposed technique with techniques using true-color images.

IV. Conclusion

The critical evaluation of the proposed method is encouraging in the direction of the low-cost design of natural user interface for home appliances, for example, radio, fans, washing machine. The strategical use of Faster RCNN (deep neural network) with SIFT algorithm, in determining the spatial and temporal location of hand (efficiency 98.7 %) and classification of hand gesture into the machine command (efficiency 99 %), make the method reliable and invariant to hand posture. The technique is simple in its design and robust in working such that the methodology can be easily integrated with any day-to-day machine at an affordable cost. The use of the Quadrant system in the design of the motion model gives the user the freedom to perform a gesture in a random path and simplifies the computation complexity in trajectory plotting for a machine. This technique is first in its type where a simple camera is used for the development of a smart and manipulative user-friendly natural interface via dynamic hand gestures. In the future, the integration of more hand postures and semantic development of trajectory and command can add new dimensions to NUIs.

References

[1] S. Yang, P. Premaratne, P., P. Vial, P. “Hand gesture recognition: An overview”. In *Proceedings of the 5th IEEE International Conference on Broadband Network & Multimedia Technology*, Guilin, China, pp. 63-69, 2013.

[2] P. K. Pisharady, M. Saerbeck. “Recent methods and databases in vision-based hand gesture recognition: A review”, *Computer Vision and Image Understanding*, (141), pp.152-65, 2015.

[3] S. Rautaray, S., A. Agrawal. “Vision based hand gesture recognition for human computer interaction: a survey”, *Artificial intelligence review*, (43), pp. 1-54, 2015.

[4] M. Van den Bergh, D. Carton, R. De Nijs, N. Mitsou, C. Landsiedel, K. Kuehnlenz, D. Wollherr, L. Van Gool, M. Buss. “Real-time 3D hand gesture interaction with a robot for understanding directions from humans”. In *Proceedings of the IEEE 2011 RO-MAN*, pp.357-362, 2011.

[5] Z. Ren, J. Yuan, J. Meng, Z. Zhang. “Robust part-based hand gesture recognition using Kinect sensor” *IEEE Transactions on Multimedia* (15), pp. 1110-1120, 2013.

[6] E. Ohn-Bar, M. M. Trivedi. “Hand gesture recognition in real time for automotive interfaces: A multimodal vision-based approach and evaluations”, *IEEE Transactions on Intelligent Transportation Systems*, 15(6), pp. 2368-2377, 2014.

[7] E. Nasr-Esfahani, N. Karimi, S. M. Soroushmehr, M. H. Jafari, M. A. Khorsandi, S. Samavi, K. Najarian. “Hand gesture recognition for contactless device control in operating rooms”, *arXiv preprint arXiv:1611.04138, 2016*.

[8] H. M. Bandara, K. S. Priyanayana, A. G. Jayasekara, D. P. Chandima, R. A. R. C. Gopura. "An Intelligent Gesture Classification Model for Domestic Wheelchair Navigation with Gesture Variance Compensation", *Applied Bionics and Biomechanics* (2020), 2020.
[9] J. Suarez, R. R. Murphy. "Hand gesture recognition with depth images: A review". In *Proceedings of the RO-MAN: The 21st IEEE International Symposium on Robot and Human Interactive Communication*, pp. 411–417, 2012.
[10] G. Marin, F. Dominio, P. Zanuttigh. "Hand gesture recognition with jointly calibrated leap motion and depth sensor". *Multimedia Tools and Applications*, *75*(22), pp. 4991-15015, 2016.
[11] D. Zhao, Y. Liu, G. Li. "Skeleton-based dynamic hand gesture recognition using 3d depth data". *Electronic Imaging*, (18), pp. 461-1, 2018.
[12] B. K. Chakraborty, D. Sarma, M.K. Bhuyan, K.F. MacDorman. 2017. "Review of constraints on vision-based gesture recognition for human–computer interaction", *IET Computer Vision*, 12(1), pp.3-15, 2017.
[13] H. Yang, L. Shao, F. Zheng, L. Wang, Z. Song . "Recent advances and trends in visual tracking: A review". *Neurocomputing*, 74(18), pp. 3823-31, 2011.
[14] Z. H. Chen, J. T. Kim, J. Liang, J. Zhang, Y. B. Yuan. "Real-time hand gesture recognition using finger segmentation", *The Scientific World Journal* (2014), 2014.
[15] G. Simion, C. David, V. Gui, C. D. Caleanu. "Fingertip-based real time tracking and gesture recognition for natural user interfaces". *Acta Polytechnica Hungarica* 13(5), pp. 189-204, 2016.
[16] P. Xu, "A real-time hand gesture recognition and human-computer interaction system." arXiv preprint arXiv:1704.07296 (2017).
[17] R. F. Pinto, C. D. Almeida A., I. C. Paula. "Static hand gesture recognition based on convolutional neural networks." *Journal of Electrical and Computer Engineering* (2019), 2019.
[18] J. Singha, A. Roy, R. H. Laskar. "Dynamic hand gesture recognition using vision-based approach for human–computer interaction". *Neural Computing and Applications*, 29(4), pp. 1129-41, 2018.
[19] W. He, T. Yamashita, H. Lu, S. Lao. "Surf tracking". In *Proceedings of the 2009 IEEE 12th International Conference on Computer Vision IEEE*, pp. 1586-1592, 2009.
[20] J. Bao, A. Song, Y. Guo. "Dynamic hand gesture recognition based on SURF tracking". In *Proceedings of the International Conference on Electric Information and Control Engineering*, pp. 338-341, 2011.
[21] Y. Yao, C. T. Li. "Real-time hand gesture recognition for uncontrolled environments using adaptive SURF tracking and hidden conditional random fields". In *Proceedings of the International Symposium on Visual Computing*, pp542-551, 2013.
[22] D. L. Dinh, J. T. Kim, T. S. Kim. "Hand gesture recognition and interface via a depth imaging sensor for smart home appliances". *Energy Procedia*, 62(62), pp.576-582, 2014.
[23] N.Ç. Kılıboz, U. Güdükbay. "A hand gesture recognition technique for human–computer interaction." *Journal of Visual Communication and Image Representation* 28, pp.97-104, 2015.
[24] D. S. Tran, N. H. Ho, H. J. Yang, E. T. Baek, S. H. Kim, G. Lee. "Real-Time Hand Gesture Spotting and Recognition Using RGB-D Camera and 3D Convolutional Neural Network", *Applied Sciences*, 10(2), pp. 722, 2020.
[25] J. Zeng, F. Wang, Y. Sun. "A natural hand gesture system for people with brachial plexus injuries". *Computing and Informatics*, 34(2), pp. 367-382, 2015.
[26] H. S. Grif, C. C. Farcas. "Mouse cursor control system based on hand gesture". *Procedia Technology* 22, pp. 657-661, 2016.
[27] S. Ren, K. He, R. Girshick, J. Sun. "Faster R-CNN: towards real-time object detection with region proposal networks." *IEEE transactions on pattern analysis and machine intelligence* 39(6) pp. 1137-1149, 2016.
[28] R. Girshick. "Fast r-cnn." In *Proceedings of the IEEE international conference on computer vision*, pp. 1440-1448, 2015.
[29] D. G. Lowe. "Distinctive image features from scale-invariant keypoints". *International journal of computer vision*, 60 (2), pp. 91-110, 2004.
[30] T. Tuytelaars, K. Mikolajczyk. "Local invariant feature detectors: a survey". *Foundations and trends® in computer graphics and vision*, 3, pp. 177-280, 2008.
[31] M. Moreira, E. Fiesler. "Neural networks with adaptive learning rate and momentum terms". *Idiap*, 1995.
[32] R. Rojas. "Fast Learning Algorithms, in Neural Networks". *Springer* 1996, pp. 183-225, 1996.
[33] P. Premaratne, S. Ajaz, M. Premaratne. " Hand gesture tracking and recognition system using Lucas–Kanade algorithms for control of consumer electronics". *Neurocomputing*, 116, pp. 242-249, 2013.
[34] N. Vandayar, T.J. McBride, K.J. Nixon. "Low Cost Hand Gesture Recognition System Design and Implementation". In *Proceedings of the Southern African Universities Power Engineering Conference/ Robotics and Mechatronics/ Pattern Recognition Association of South Africa, IEEE*, pp. 217-222, 2019.
[35] A.G. Cruz Bautista, J. J. González-Barbosa, J. B. Hurtado-Ramos, F. J. Ornelas-Rodriguez, E. A. González-Barbosa. "Hand features extractor using hand contour–a case study", *Automatika*, 61(1), pp. 99-108, 2020.
[36] E. Thabet, F. Khalid, P. Suhaiza Sulaiman, R. Yaakob. "Low cost skin segmentation scheme in videos using two alternative methods for dynamic hand gesture detection method", *Advances in Multimedia*, *2017*.

Author Biographies

Richa Golash B. E. in E & C from Dr K.N.M.I.E.T., India in 2000, M. Tech in 2008, currently pursuing Ph. D. from R.G.P.V. India. She has teaching experience of 15 years and have deep interest in the field of artificial intelligence. Published many research papers in various International/ National Journals/ Conferences.

Yogendra Kumar Jain: B. E. in EC & I Engg. from S. A. T. I., M. P., India in 1991, M. E. in Digital Tech. & Instru. from SGSITS, India. Awarded Ph. D. from R.G.P.V. India in 2010. He has teaching experience of 28 Years, published more than 150 research paper in various International/ National Journals/ Conferences. He is a reviewer and an active editorial board member of many reputed journals .