Real-Time Hand Gesture Recognition System

Aarti[1,†], Swathi Gowroju[2,*,†], Raju Pal[3,†], Vaddiraju Swathi[2,†] and Sirisha Yerraboina[4,†]

[1]*Lovely Professional University, Punjab, INDIA*
[2]*Sreyas Institute of Engineering and Technology, Hyderabad, Telangana, INDIA*
[3]*Jaypee Institute of Information Technology, Noida, Uttar Pradesh, INDIA*
[4]*Matrusri Engineering College, Hyderabad, Telangana, INDIA*

Abstract

A machine and a person can interact through hand gestures by using a hand gesture identification device. Real Time Hand Gesture Recognition (RTHGR) is discussed in this work in order to carry out system control actions as intended. With the help of this application, the user's hand gestures may be detected by the webcam and basic actions can be taken as a result. The user must make a distinct gesture. The webcam records this, recognizing the gesture and carrying out the action in accordance with a list of recognized gestures. This process requires a binary threshold value to recognize gestures. A neural network is employed in the proposed classification process. The effectiveness of this technique for operating various systems will be assessed, and other hand recognition techniques will be compared.

Keywords

Gesture recognition, CNN, YOLO, Region of Interest, deep learning, object recognition

1. Introduction

The world is going through a technological revolution. The rapid advancement in the technological aspect of human beings is growing faster than ever which has allowed the various computer systems to indulge in our everyday lives as an integral part of our daily activities. Various computer systems have different methods of interaction with the users. This is known as HCI in technical term. It is a very important aspect of intractability, usefulness, practicality, and overall user experience on a computer system. In the past few years, the user experience is focused more than anything on a system made to be used by humans. This is because the effectiveness of a system is greatly measured based on how a system interacts with the user to make the overall experience of using a system easier and more wonderful. In the last decade, hardware like keyboard, mouse and touch-screen have been crucial in how people engage with technology. However, new forms of engagement tools have been created as a result of the quick advancement of technology. In the realm of HCI, technologies like thought processing, gesture recognition, and speech recognition have advanced significantly. In our proposed system, gesture recognition is one of these that is covered. In this, hand gestures are utilized as communication between humans and electronic devices. It differs significantly from

ACI'23: Workshop on Advances in Computational Intelligence at ICAIDS 2023, December 29-30, 2023, Hyderabad, India

[*]Corresponding author.

[†]These authors contributed equally.

aarti.1208@gmail.com (Aarti); swathigowroju@sreyas.ac.in (S. Gowroju); raju3131.pal@gmail.com (R. Pal)

conventional hardware-based techniques, which are capable of achieving human-computer interaction on a totally other level. The subject of computer vision and image processing has been substantially altered by convolutional neural networks (CNNs), a significant advancement in artificial intelligence and machine learning. These neural networks were created expressly to tackle visual perception, one of the most difficult and inherently human jobs. CNNs have advanced standard machine learning techniques by imitating the hierarchical and feature-driven nature of how we, as humans, perceive and recognize patterns in the visual environment. CNNs were inspired by the complex workings of the human visual system. By doing this, they have unlocked astonishing ability for machines to understand, categorize, and extract valuable information from photos in addition to allowing them to ”see” images. CNNs operate on small, overlapping regions of the image known as receptive fields, which allow the network to capture local patterns and gradually build a rich hierarchical representation of the input data. CNNs are a class of deep neural networks distinguished by their unique architecture, which includes convolutional layers and pooling layers. A wide range of fields have been significantly impacted by the introduction of CNNs. In addition, they have found use in a variety of fields, including object identification, facial recognition, autonomous cars, medical picture analysis, and more.

2. Literature Survey

Since the subject of hand gesture recognition is expanding quickly, as many implementations employing both deep learning and machine learning algorithms that attempt to identify a gesture that is exhibited by a human using his/her hand. The study [1, 2] showed the common and upcoming machine learning designs, CNN, achieved faster rates of successfully perceiving components at a negligible computational cost. The suggested strategy focused primarily on instances of movements that existed in pictures with two sets such as with hand gestures and without hand gestures placement and followed instances of hand obstruction with 24 movements. It used a segmentation algorithm and back propagation algorithm to prepare the multi-layer propagation, and for the back propagation going backwards from nodes that produce to input nodes in order to check for faults to have an impact, sorting. Among these approaches, Hidden Markov Models (HMM) is a well-liked technique that is employed by a number of other detecting applications. The proposed system in this article refers to checks and operates with all of the numerous detection versions that are frequently employed by an application that we have seen by studying and examining other papers such as image, video, and webcam are all addressed. According to a generic document produced by Francois et al. [3], the posture detection application he refers to employs video and the HMM to do so. The three aforementioned approaches all identify these features, but whether they are based on CNNs, RNNs, or some other technique, the primary issue with all of them is that they all employ fitting techniques, which all make reference to the bounding box [4] that was covered in this study. The output of what image is being presented is determined by the confidence value that is the highest, which is derived from the bounding box that represents the data that is detected. Certain additional tools and methods connected to segmentation, general localization, and even the union of other different areas aid in the accomplishment of the tasks of detecting and recognizing. Fuzzy based human behavior recognition model was developed based on the body gestures [5]. Rahim et al. [6] analysis uses the conversion of a signed language word's gesture into text. The skin mask segmentation was used by the authors of this work used basic CNN model to extract features. The support vector machine algorithm is a kind of supervised learning method used to address regression and classification problems in machine learning used to classify the signs' movements with a 95.28% accuracy using a dataset of 10 movements from one hand and 8 from both the hands [7]. Mambou et al. [8] analysis of nighttime both indoors and outdoors included hand gestures connected to sexual assault. The YOLO CNN architecture was used to create the gesture recognition system. This architecture extracted hand motions and then classified bounding box images to provide the assault alert. MOving object classification was done using eigen faces and optical flow approaches [9]. To decode gestures or finger-spelling from movies that express multiple letters are signed in a series to form meaningful words by identifying finger spellings in uncut sign language footage is a hard process, Ashiquzzaman et al. [10] presented the lightweight spatial prism pooling (SPP) using a CNN model. The model performed 3 times faster than conventional models and required 65% fewer parameters than conventional classifiers. A lightweight semantic segmentation, Fast and Acurate Semantic Segmentation Dilated-Net network was used by Benitez-Garcia et al. [11] in place of Temporal Segment Networks (TSN), It is predicated on the notion of modeling long-range temporal structures and Spatiotemporal Shift Modules (SSM). On a dataset of thirteen gestures aimed at real-time interaction with touch less screens, they proved the effectiveness of the idea. There are several other CNNs [12, 13, 14, 2, 15] that implemented mark-based prediction accurately up to 98% for various biometric applications. Most of the publications [16, 17] concentrate on the data gathering, surroundings, and hand gesture representation three essential components of the "vision-based hand-gesture" identification system. We have also evaluated the vision-based recognition of hand movements system's performance in terms of recognizing precision. The prediction accuracy for the signer dependent, CNN was used to train a total of 21 ISL static alphabets, yielding verification and testing accuracy of 97.34% and training accuracy of 98.50%.. On the other hand, the signer independent's claimed identification accuracy varies from 50-90%, with a standard recognition accuracy of 78.2%, according to the studies that were chosen. Musa et al. developed some models to identify and trace the suspicious activities based on body movements [18, 19].

3. Proposed System

In this investigation of CNNs, we will enlarge on their structural elements, the guiding principles for their success, and the various applications that make use of their extraordinary capabilities. We'll show how these networks have changed computer vision and ushered in a new era of artificial intelligence by enabling machines to understand and interact with the visual environment. CNNs have shown to be quite successful at recognizing hand gestures. In order to effectively categorize and interpret these motions, CNNs are employed in the context of hand gesture recognition to automatically extract features from pictures or video frames including hand gestures. Fig. 1 shows various steps of proposed hand gesture recognition system.

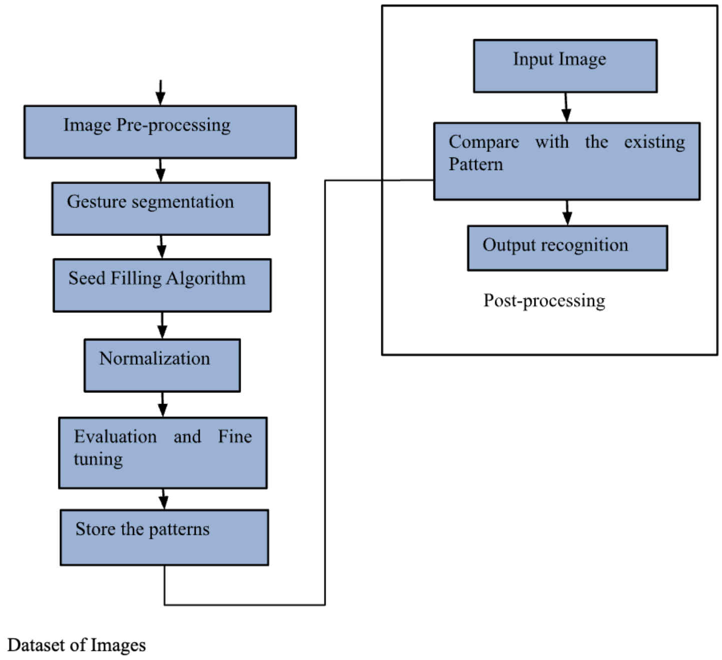

Figure 1: Proposed System

3.1. Data Gathering and Pre-processing:

A dataset of hand gesture photos or video frames is normally gathered in order to train a CNN for hand gesture identification. This dataset should contain a variety of hand gestures made by various people in various situations. Preparing the data is a crucial step before training and testing. Each image was duplicated for reading and training, for two hands, by flipping it horizontally, occasionally taking the corresponding image from both hands for making the set more precise. Proposed system uses YOLO architecture about 230+ images of the dataset, where 100 images were utilised for the testing by promoting it to 2-fold. 15 more pictures were also captured and labelled for the testing set. Data pre-processing is essential before post-processing so that we can identify the kind of data we collected and which parts will be useful for developing, evaluating, and enhancing accuracy.

A represents X-Value, B represents Y- Value, C represents Width, and D represents Height. Table 1 illustration depicts how these files would appear when we label our dataset in order to train it on the desired model.

Five different traits, each with their own significance, are included in each line. The class ID is the first item on the left, are the coordinates that define the labelled box around the gesture. It includes the x-axis and y-axis values, specifying the position of the top-left corner and the bottom-right corner of the bounding box.

Table 1
Data acquisition from data file of YOLO

Classification Id	A	B	C	D
0	0.531771	0.490234	0.571875	0.794531
1	0.498437	0.533203	0.571875	0.905469
2	0.523438	0.579297	0.613542	0.819531
3	0.526563	0.564453	0.473958	0.819531
4	0.498611	0.587891	0.977778	0.792969

3.2. Gesture Segmentation:

The challenging nature of gesture training increases as a result of the fact that gesture data is recorded in various places, under various lighting conditions, and at various times of day. The RGB color space data is transformed into the YCbCr color space in the color image pipeline. This conversion allows the separation of chroma (Cr) and brightness (Cb), effectively mitigating interference from brightness characteristics.

$$\left[yC_bC_\gamma\right] = [0.210.71230.0692 - 0.1146 - 0.37850.490.48 - 0.4653 - 0.0456][RGB]$$

$$[RGB] = \left[\begin{array}{ll} 101.57381 & -0.1873 - 0.46511.84920 \end{array}\right] \tag{1}$$

Using Eq 1 the Cr channel is extracted. The data picture is subjected to Gaussian filtering, and the 2D Gaussian distribution is used to limit the impact of Gaussian noise is characterized by a bell-shaped curve, and Gaussian noise is essentially random noise with a mean of zero using Eq 2.

$$f(i,j) = f(\mathbb{1})f(j) = \frac{1}{\sqrt{2\Pi\sigma}}\mathfrak{e}^{e^{\frac{-(\mathbb{1}-u_i)^2}{2}}\sigma_i^2}\frac{1}{\sqrt{2\Pi\sigma}}\mathfrak{e}^{e^{-\frac{-(t-u_j)^2}{2}}\sigma_j^2} \tag{2}$$

The Eq 2 belong to the bivariate Gaussian distribution family, with i and j representing the two dimensions, and i and j representing the two dimensions, and ui, uj, σi, and σj representing the mean and standard deviation parameters for each dimension. The difference between the Gaussian template coefficients will be lower and the smoothing impact on the image will be more noticeable as the value increases.

3.3. Seed Filling Algorithm:

To separate touching objects in the image the proposed system uses the mark-based algorithm that assigns pixels to regions based on proximity to markers, resulting in segmented regions delineated by watershed boundaries is used to segment the data images for the image samples processed by skin colour detection. The algorithm connects neighbouring pixels with similar grey values, forming contours for image segmentation. This approach is especially effective for images with noise and irregular gradients, simplifying the segmentation process by highlighting distinct intensity patterns in the image. Since the standing level of the component that is connected can be raised like a dam, the watershed algorithm based on mark may stop the local smaller edges are merged and inseparable. Through the supervision of the mark-based watershed algorithm, the gesture features may be efficiently segmented. The 8-connected edge filling method is used to sporadic gesture portion once the gesture features have been accurately segmented. An enhancement on the four-connected filling process is the 8 edge connected seed filling algorithm. Unlike the four-connected filling method, which begins the process at the beginning, the 8-connected seed filling technique spreads in eight directions, speeding up the process. at the centre of injection in the area and expands in each direction to cover all of the area's pixels. The algorithm used to obtain the sign feature data has eight connected seeds.

3.4. Normalization and Gesture Labelling:

After the image data has been filled in and segmented, the scale normalisation operation may guarantee the reliability of the feature extraction and the processing of labelling the data with gestures clear features for training the model efficiently. In the proposed method, the segmented and filled picture data are normalised to 128x128, which effectively increases the model's performance, speed and accuracy during training and prevents gradient explosion.

3.5. Evaluation and Fine-Tuning:

Following training, the model is evaluated on a different validation or test dataset to gauge its performance in terms of accuracy and generalization. To attain the appropriate level of accuracy, the model may need to be fine-tuned and its hyper-parameters may need to be adjusted.

3.6. Post-processing:

Post-processing techniques may be used to improve the precision of gesture detection in real-world scenarios or to smooth predictions over time.

4. Results & Analysis

We tested on computer with an AMD Ryzen 5 processor and a 64-bit version of Windows 10, our experiment is run using OpenCV and PIL installed for image processing, and Anaconda is installed to create an interactive interface, were used to implement the identification and classification operations.

4.1. Experimental Analysis:

The data set is collected from publicly available sources, and each gesture data gathers 250 image data. The case study's trained model is capable of identifying ten distinct movement types, ranging from 0 to 9. First, a total of 1980 data samples representing twenty types of gestures are gathered for each class of 10 individuals over a range of time periods. The training outcome for gestures 3, 4, and 5 is poor. The training part can be significantly increased by including sample data. As a result, 100 gestures are added to each category of gestures, which considerably raises the success rate of recognition.

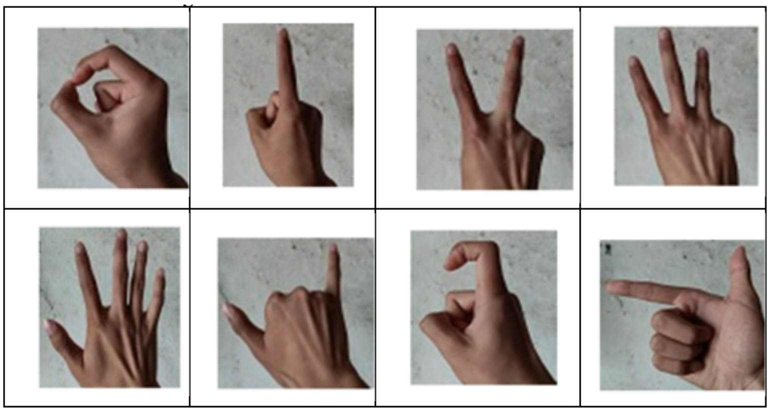

Figure 2: Sample images from dataset

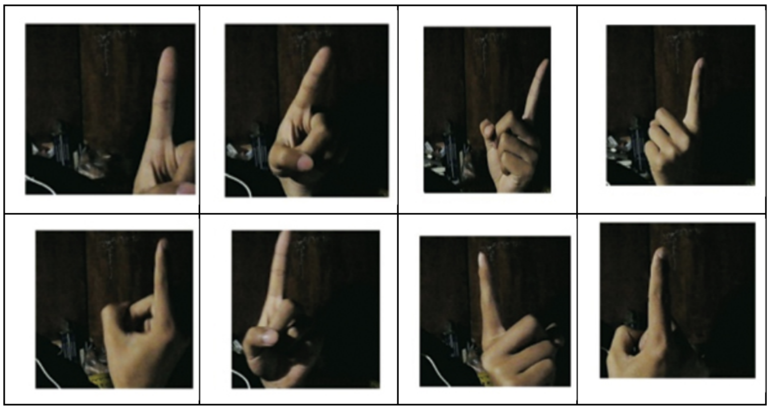

Figure 3: Sample data for same gesture

A comprehensive data set requirement is necessary to enhance training effect. The position on the photograph will differ for the same motion when performed by different people, at different angles, and with variable lighting conditions. Fig. 3 displays various data for the same move. In the detection rate test, 150 test specimens were collected for quantifying across various time periods; 95 of the samples had been successfully identified, 7 of the samples experienced recognition mistakes, and 1 sample did not exhibit any gesture. The test's outcomes in the very dark environment weren't the best; there were four times as many mistakes in the prediction of gestures 2 and 3. This is due to the fact that the predicted gesture features cannot be distinguished in setting and because gestures 4 and 5 are too similar to one another. In a typical context, the model generated by the data pre-processing method can attain an optimal prediction rate by improving the gesture characteristic. Data from four groups of samples is collected and divided into the model's robustness and stability tests. The analysis result is shown in Fig. 4.

We have introduced a real-time hand gesture recognition system in this research that makes

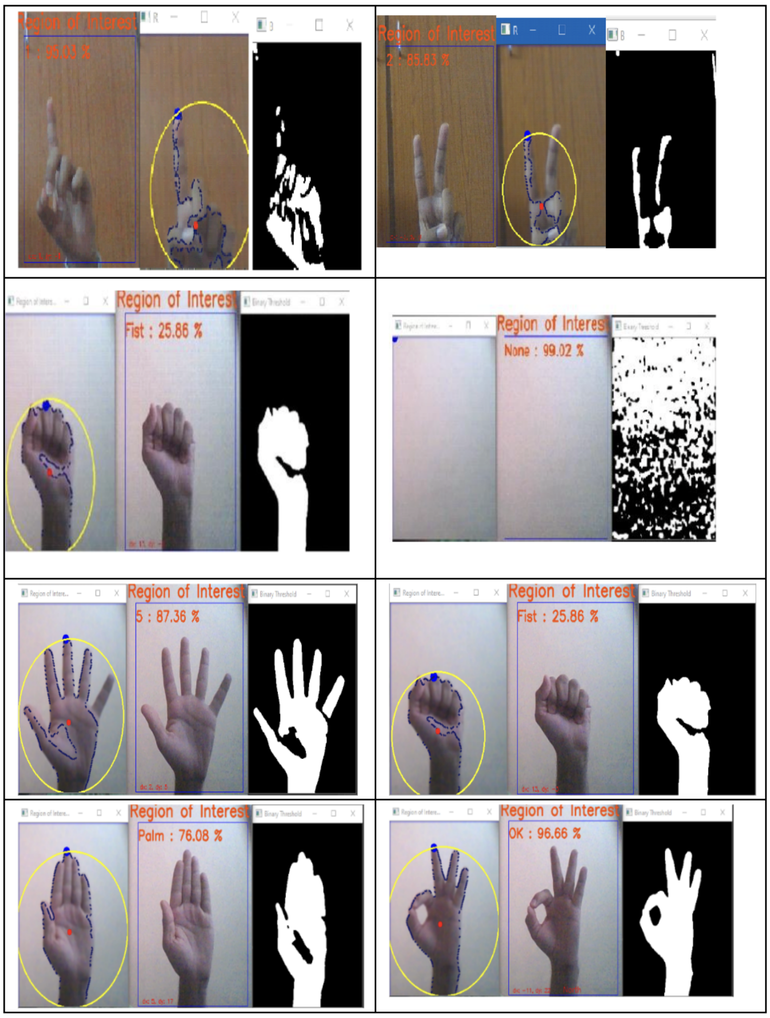

Figure 4: Evaluation results of proposed system

use of cutting-edge computer vision and machine learning algorithms to precisely interpret and categorize hand gestures in practical settings. Our study has significantly advanced the field of gesture recognition in a number of ways.

4.2. Robustness and Accuracy

We have shown that our system achieves accuracy in recognizing a wide variety of hand gestures through thorough experimentation and review. While CNNs excel in extracting spatial features from grid-like data such as images, RNNs are well-suited for sequential data with temporal dependencies. The choice between them depends on the nature of the task and the characteristics of the input data. among other deep learning models, have considerably increased the system's capacity to handle complicated and dynamic movements. The main goal of our research was to create a system that could process data in real-time. Our system can now process and recognize gestures in real time, enabling technologies or methodologies likely play a crucial role in improving user experiences and enhancing interactions in these areas. This indicates that we have successfully accomplished our goal.

4.3. Comparison of each group

In a typical context, the model produced by the data pre-processing approach can reach an optimal recognition rate by improving the gesture feature. Four sample groups of test data, each with 30 test data graphs, are randomly chosen from the test data set in order to assess the model's stability and robustness. Each data group was recognized and put to the test. The model exhibits strong resilience and stability after testing. Table 2 presents the comparing findings.

Table 2
Accuracy comparison

Group No	Accuracy Rate (%) of recognition
The 1st Category	95.62
The 2nd Category	94.47
The 3rd Category	96.00
The 4th Category	98.12

The model's convergence rate is accelerated since dropout is used to prevent over-fitting. After numerous tries, increasing the batch value in the training process can lower the number of epochs and accelerate training. It could not be able to effectively collect the data characteristics due to the decreased number of repetitions, which would lower the prediction rate. When the batch size reaches 35 following testing, the training time and model constancy are well-aligned. The training is stopped early after the saturation point. The loss curve comparison chart for various batch values is shown in Fig. 5.

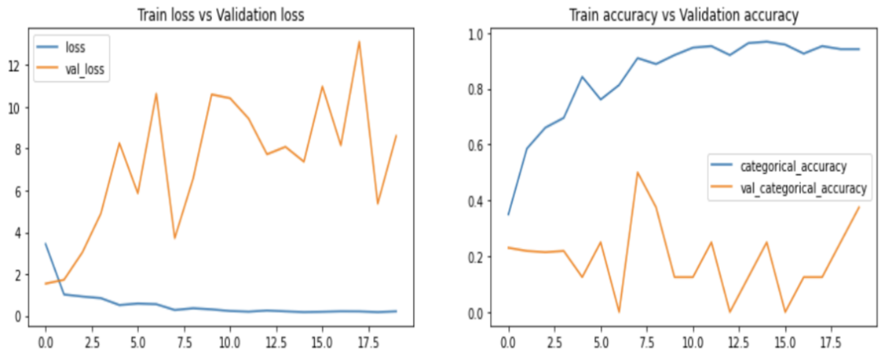

Figure 5: Loss and accuracy graphs of training and testing

Three people are chosen at random to participate in the laboratory testing to confirm the acceptance of prediction of gestures on various participants in proposed system. Each of the 10 motions must be tested 500 times, with each tester testing each position nearly 50 times, by cross folding into two groups and testing each gesture. If testing is needed during the day, each group must take it 15 times, and if it is needed at night, each group must take it 250 times. The three testers' respective recognition rates for the ten gestures are 94.4%, 94.6%, and 93.8%.

5. Conclusion

In this study, the gesture data is pre-processed using skin colour recognition, marker-based watershed algorithms, and seed filling algorithms. in order to generate the gesture data with clear gesture features. Proposed system accurately performed on the test data and then achieves 98.66% under conditions of ordinary light by using the training of 10 different types of gesture data following YOLO convolution neural network pre-processing. The pre-processing technique applied in this study effectively mitigates the influence of the surrounding background on gesture recognition and detection. Notably, its implementation does not necessitate additional training or detection time. Additionally, the post-processing step would depend on the specific requirements of our gesture recognition task.

References

[1] R. Khan, N. A. Zaman, Hand gesture recognition: a literature review, International journal of artificial Intelligence & Applications 3 (2012).

[2] A. Swathi, Aarti, S. Kumar, A smart application to detect pupil for small dataset with low illumination, Innov. Syst. Softw. Eng. 17 (2021) 29–43.

[3] A. Caputo, A. Giachetti, S. Soso, D. Pintani, A. D'Eusanio, S. Pini, G. Borghi, A. Simoni, R. Vezzani, R. Cucchiara, A. Ranieri, F. Giannini, K. Lupinetti, M. Monti, M. Maghoumi, J. J. LaViola, Jr, M.-Q. Le, H.-D. Nguyen, M.-T. Tran, SHREC 2021: Track on skeleton-based hand gesture recognition in the wild (2021). arXiv:2106.10980.

[4] M. Chmurski, G. Mauro, A. Santra, M. Zubert, G. Dagasan, Highly-optimized radar-based gesture recognition system with depthwise expansion module, Sensors (Basel) 21 (2021) 7298.

[5] P. Agrawal, V. Madaan, N. Kundu, D. Sethi, S. K. Singh, X-hubis: A fuzzy rule based human behaviour identification system based on body gestures, Indian Journal of Science and Technology (2016) 1–6.

[6] M. A. Rahim, M. R. Islam, J. Shin, Non-touch sign word recognition based on dynamic hand gesture using hybrid segmentation and CNN feature fusion, Appl. Sci. (Basel) 9 (2019) 3790.

[7] N. Mohamed, M. B. Mustafa, N. Jomhari, A review of the hand gesture recognition system: Current progress and future directions, IEEE Access 9 (2021) 157422–157436.

[8] Mambou, Krejcar, Maresova, Selamat, Kuca, Novel hand gesture alert system, Appl. Sci. (Basel) 9 (2019) 3419.

[9] P. Agrawal, R. Kaur, V. Madaan, M. S. Babu, D. Sethi, Moving object detection and recognition using optical flow and eigen face using low resolution video, Recent Advances in

Computer Science and Communications (Formerly: Recent Patents on Computer Science) 13 (2020) 1180–1187.

[10] A. Ashiquzzaman, H. Lee, K. Kim, H.-Y. Kim, J. Park, J. Kim, Compact spatial pyramid pooling deep convolutional neural network based hand gestures decoder, Appl. Sci. (Basel) 10 (2020) 7898.

[11] M. Gibran, Y. Haris, N. Tsuda, Continuous finger gesture spotting and recognition based on similarities between start and end frames, IEEE Transactions on Intelligent Transportation Systems 23 (2020) 296–307.

[12] S. Gowroju, S. Kumar, Aarti, A. Ghimire, Deep neural network for accurate age group prediction through pupil using the optimized UNet model, Math. Probl. Eng. 2022 (2022) 1–24.

[13] S. Gowroju, Aarti, S. Kumar, Robust pupil segmentation using UNET and morphological image processing, in: 2021 International Mobile, Intelligent, and Ubiquitous Computing Conference (MIUCC), IEEE, 2021.

[14] S. Gowroju, S. Kumar, Aarti, A. Ghimire, Deep neural network for accurate age group prediction through pupil using the optimized UNet model, Math. Probl. Eng. 2022 (2022) 1–24.

[15] S. Gowroju, Aarti, S. Kumar, Review on secure traditional and machine learning algorithms for age prediction using IRIS image, Multimed. Tools Appl. (2022).

[16] N. Mohamed, M. B. Mustafa, N. Jomhari, A review of the hand gesture recognition system: Current progress and future directions, IEEE Access 9 (2021) 157422–157436.

[17] M. Chmurski, G. Mauro, A. Santra, M. Zubert, G. Dagasan, Highly-optimized radar-based gesture recognition system with depthwise expansion module, Sensors (Basel) 21 (2021) 7298.

[18] A. S. Ben-Musa, S. K. Singh, P. Agrawal, Object detection and recognition in cluttered scene using harris corner detection, in: 2014 International Conference on Control, Instrumentation, Communication and Computational Technologies (ICCICCT), 2014, pp. 181–184. doi:10.1109/ICCICCT.2014.6992953.

[19] A. S. B. Musa, S. K. Singh, P. Agrawal, Suspicious human activity recognition for video surveillance system, in: IEEE proceedings of 2014 international conference on control, instrumentation, communication and computational technologies, 2014, pp. 214–218.

[20] A. Swathi, S. Kumar, V. Subbamma., S. Rani, A. Jain, R. Kumar, Emotion classification using feature extraction of facial expression, in: 2022 2nd International Conference on Technological Advancements in Computational Sciences (ICTACS), IEEE, 2022.