Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

VIRTUALLY CONTROLLING COMPUTERS USING HAND GESTURE AND VOICE COMMANDS

Mayank Sharma, Akilesh G,
Vishwaa VS, Sharon Femi P, Kala A
UG Scholar, Department Of Information Technology,
Sri Venkateswara College Of Engineering
Department Of Information Technology, Sri Venkateswara College Of Engineering

Abstract: *In recent years, using a computer in a virtual environment has attracted the interest of researchers. This paper presents an application that allows gadgets to be more interactive and functional with little or no physical interaction. To interface with the system, the program provides voice gesture recognition and hand gesture recognition functions. The interactive system begins with the user being greeted by a voice assistant named "Proton." The assistant uses the user's input to accomplish activities like as search, file navigation, date and time, and to start and stop the hand gesture recognition system. A camera records a live video stream in a Hand Gesture Recognition system, from which the input is collected via an interface. The system has been programmed to recognize a variety of hand movements that may be used to do operations such as right-clicking, left-clicking, opening a file, and so on. Our system detects the coordinates on the hand and recognizes the form and motion when the user holds their hand in front of the camera. This is accomplished through the use of the blazing palm detector model and the hand landmark model. In previous systems, data gloves or markers were used to input data into the system. There are no such limitations in this system. In front of the camera, the user may intuitively make hand motions. Hand gesture recognition systems that are totally reliable are currently being researched and developed. The created system acts as the basis for future work that may be expanded.*

Keywords: **Human Computer Interaction, Hand Gesture**

1. Introduction

Human motion gesture detection is the most difficult study path in computer vision, and it is widely employed in domains such as human-computer interaction, intelligent monitoring, virtual reality, human behaviour analysis, and others. As we all know, hand gesture detection using vision-based technologies is an essential aspect of human-computer interaction (HCI). In recent decades, the keyboard and mouse have become increasingly important in human-computer interaction. However, new sorts of HCI solutions have been necessary as a result of the fast growth of technology and software. In the subject of HCI, technologies like speech recognition and gesture recognition garner a lot of attention. A gesture is a visual representation of physical action or emotional expression. It consists of both body and hand gestures. There are two types of gestures: static gestures and dynamic gestures. A sign is denoted by the posture of the body or a hand gesture in the case of the former. The latter sends signals through the movement of the body or the hand. Gestures can be used to communicate between humans and computers. It differs significantly from traditional hardware-based techniques in that it allows for human-computer interaction via gesture recognition. The user intent is determined through gesture recognition, which recognises the gesture or movement of the body or body components. Many academics have worked for decades to develop hand motion detection technologies. Many applications, such as sign language recognition, augmented reality (virtual reality), sign language interpreters for the impaired, and robot

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

control, rely heavily on hand gesture recognition.

1 Corresponding Author

The rest of the paper is systematized as follows: Chapter 2 deals with related works, Chapter 3 deals with the existing work, Chapter 4 deals with the proposed system; chapter 5 deals with the methodology and chapter 6 shows the implementation. The conclusion is provided in Chapter 7.

2. Related works

Al-Hammadi et al. developed a Deep Learning-Based Approach for Sign Language Gesture Recognition with Efficient Hand Gesture Representation [1]. Their paper contributed research on optimizing the level of C3D architecture knowledge transfer between human activity recognition and hand gesture recognition and presenting a hand gesture recognition system based on an optimized C3D architecture. Their proposed system uses local and global configurations efficiently with more attention to the hand region, presenting a novel method for hand segmentation based on the open pose framework and optimizing two architectures for local features aggregation. In their paper, a novel system was proposed for dynamic hand gesture recognition using multiple deep learning architectures for hand segmentation, local and global feature representations, and sequence feature globalization and recognition. Their proposed system was evaluated on a very challenging dataset, which consists of 40 dynamic hand gestures performed by 40 subjects in an uncontrolled environment.

Alzahrani et al. developed a comprehensive evaluation of skeleton features based fall detection from Microsoft Kinect [2]. They experimentally evaluated their paper on skeleton features-based fall detection by comparing fall detection performance for different combinations of skeleton features used in previous related works. They determined the skeleton features that best distinguished fall froth non-fall frames, and the best performing classifier. They followed the classical five steps of supervised machine learning, they collected a learning data composed of 42 fall and 37 non-fall videos from FallFree, they extracted and preprocessed the skeleton data of the training set and they extracted each possible skeleton feature. Finally they evaluated all extracted and selected features using two main experiments, one of them based on neighborhood component analysis (NCA).

Plouffe *et al.* developed Static and Dynamic Hand Gesture Recognition in Depth Data Using Dynamic Time Warping [11]. They discussed the development of natural gesture user interface that tracks and recognizes hand gestures collected by a Kinect sensor. Their methodology used k-curvature algorithm which was employed to locate the fingertips over the contour, and dynamic time warping was used to select gesture candidates and also to recognize gestures by comparing an observed gesture with a series of prerecorded reference gestures. Two possible applications of their work were for interpretation of sign digits and gestures for a friendlier human machine interaction, the other one for the natural control of a software interface.

Banerjee *et al.* developed Mouse Control using a Web Camera based on Color Detection [3]. In this paper they presented an approach for Human computer Interaction

(HCI), where they have tried to control the mouse cursor movement and click events of the mouse using hand gestures. Hand gestures were acquired using a camera based on color detection technique.

Duan et al. uses a deep convolutional stacked hourglass network to accurately extract the location of key joint points on the image [5]. The generation and identification part of the network is designed to encode the first hierarchy (parent) and the second hierarchy (child) and show the spatial relationship of human body parts. The generator and the discriminator are designed as two parts in the network, and they are connected together in order to encode the possible relationship of appearance and, at the same time, the possibility of the existence of human body parts and the relationship between each part of the body and its parental part coding. In the image, the key nodes of the human body model and the general body posture can be identified more accurately.

Franco *et al.* developed a multimodal approach for human activity recognition based on skeleton and RGB data [6]. Their proposal focuses on the use of vision-based techniques which guarantee a higher degree of unobtrusiveness with respect to sensor-based approaches. In their work the potentialities of the Kinect sensor were fully exploited to design a robust approach for activity recognition combining the analysis of skeleton and RGB data streams.

Xie et al. developed Accelerometer-Based Hand Gesture Recognition by Neural Network and Similarity Matching Accelerometer-based pen-type sensing device and a user-independent hand gesture recognition was their algorithm [12]. Gestures in their system were divided into two types, the basic gesture and the complex gesture, which were represented as a basic gesture sequence. A dictionary of 24 gestures, including 8 basic gestures and 16 complex gestures, were defined. An effective segmentation algorithm was developed to identify individual basic gesture motion intervals automatically. Users can hold the device to perform hand gestures with their preferred handheld styles. Through segmentation, each complex gesture is segmented into several basic gestures. Based on the kinematics characteristics of the basic gesture, 25 features were extracted to train the feed forward neural network model. The input gestures were classified directly by the feed forward neural network classifier.

3. Existing System

Many other ways of identification have been presented. Al-Hammadi et al. [1] employed a system that uses local and global configurations efficiently with more attention. Alzahrani et al. [2] evaluated the skeleton features that distinguish fall from non-fall frames. Banerjee et al. [3] used colour detection technique to detect hand gestures. Cui et al. and Franco et al. [4,6] also employed Kinect to acquire depth information and created a human position detection system that was especially built for sitting and standing postures. [5] discovered the global optimum features from a variety of characteristics, including Fourier descriptors, form context, edges, and gradients, to complete the back-projection procedure from features to three-dimensional poses rapidly and reliably. The literature [6] employs Markov random field (MRF) to segment point cloud data including the human body into four parts: head, torso, limbs, and background, as well as part identification to recognise the body's position. Jiang et al. [7] used skeleton

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

algorithm and CNN to reduce the impact of shooting angle and improve the accuracy of gesture recognition in complex environment. To get location information for each region of the human body, [8] proposed learning the relevant target contour model from the segmented picture and then using the boost classifier to discover the contour of the target in the image using the boost classifier. Li et al. [9] employed the canny operator in conjunction with pixel depth information to extract edge features from the picture, determining the head location of the person in the image by distance transformation and model matching, and positioning the human body according to the prior human body proportion. Nadeem et al. [10] created a 3D grid model of the human body in order to locate places of interest connected to the grid's measurement extreme value, and then utilised these points to detect the human body's head, hands, and feet. Plouffe et al. [11] collects data from kinetic sensor to track and recognize hand gestures in real time. Xie et al. [12] used accelerometer-based pen-type device and a user-independent hand gesture recognition system. To achieve the goal of human posture estimation, Zhang et al. [13] employed the iterated closest point (ICP) approach to monitor the initialised human skeleton. Zhang et al. [14] employed the histogram of gradient directions and trained several local linear regressions.

4. Proposed System

The camera gathers the input initially, as illustrated in Figure 1, and then runs it through two algorithms: the Palm Detection Model and the Hand Landmark Model. The algorithm identifies movements and executes various actions based on the gestures we make. The microphone input identifies our voice and conducts various tasks based on the input.

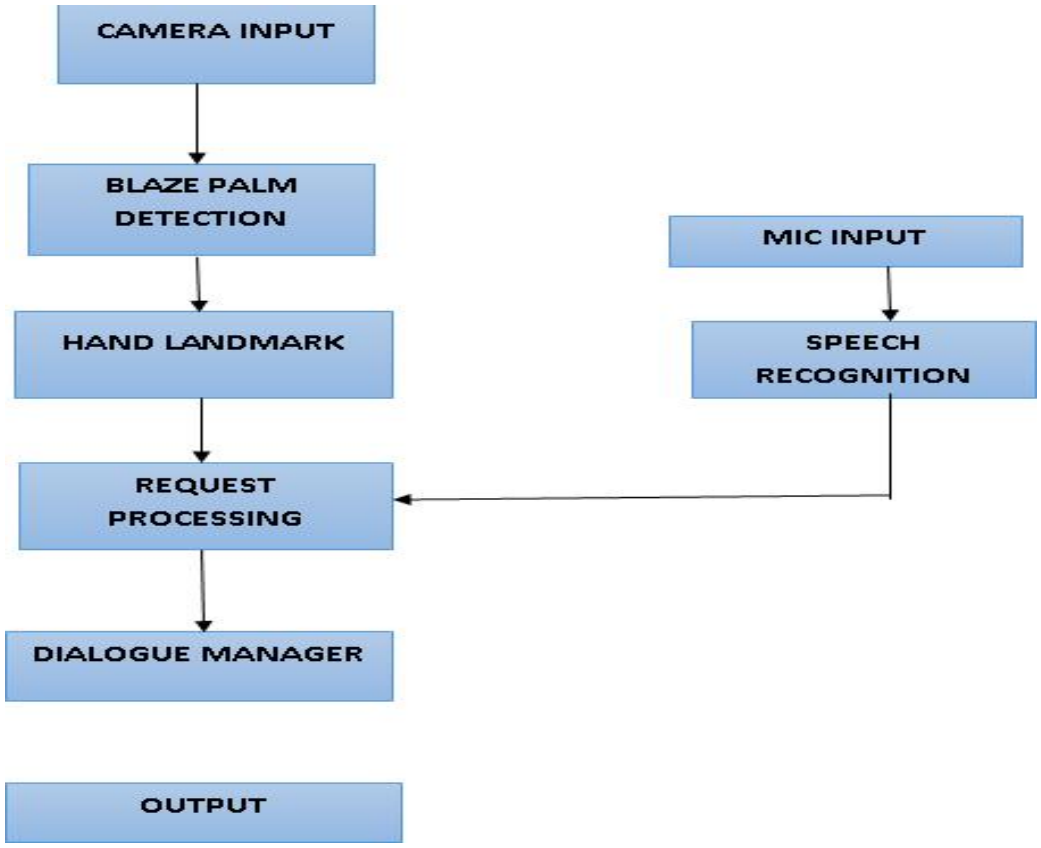

Figure 1. Gesture Recognition Architecture diagram

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

Methodology

Input Module

Input module includes the Camera input and the Microphone input. The camera captures the input frame by frame and is processed by palm recognition model and the flame palm model. The microphone detects the voice and turns it to text using the speech recognizer.

Blaze Palm Detector module

A one shot detector model is used to identify initial hand placements, and it is ideal for mobile real-time applications. Hand identification is a tough problem operating with a broad range of hand sizes and a large scale span (20x). Faces contain strong contrast patterns around the eyes and mouth, while hands lack similar traits, making it more difficult to consistently distinguish them based on their visual features alone. To begin with, a palm detector is trained since estimating bounding boxes of rigid objects like palms and fists is considerably easier than distinguishing hands with articulated fingers. Then, an encoder-decoder feature extractor similar to FPN is used for a greater scene-context awareness, even for little objects. Finally, concentration loss during training is minimised to support a large number of anchors resulting from the high scale variance. A high-level palm detector concept is shown in Figure 2. In palm detection, this approach has so far obtained an average precision of 95.7 %.

Figure 2. Palm Detection Module

Hand Landmark model

The hand landmark model employs regression to provide exact landmark localization of 21 3D coordinates inside the observed hand areas, as illustrated in Figure 3. The model produces a consistent internal hand posture representation even with partially visible hands and self-occlusions. The model generates three outputs. Three outputs share a feature extractor in the model. Corresponding datasets in the same hue are used to train each head. There are 21 hand landmarks with x, y, and relative depth. The presence of a hand in the input picture is indicated by a hand flag. Left or right hand is a binary categorization of handedness.

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

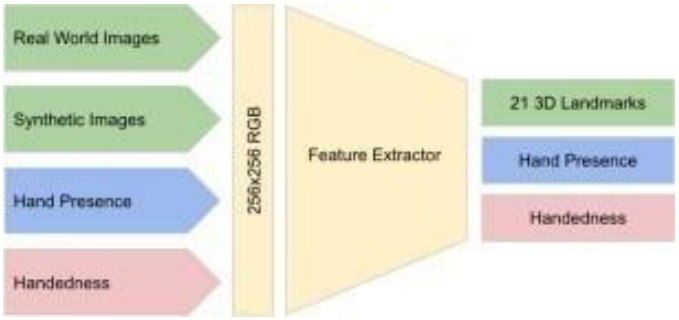

Figure 3. Hand Landmark Module

The real-world and synthetic datasets are used to learn the 2D coordinates, while synthetic pictures are used to learn the relative depth with respect to the wrist point. Furthermore, palms may be simulated using square bounding boxes that ignore other aspect ratios, resulting in a reduction of 3-5 anchors. Second, even for little objects, an encoder-decoder feature extractor is employed for larger picture context awareness. Finally, to support a large number of anchors coming from the high scale variance, reduce focus loss during training. A second model output evaluates the chance of an adequately aligned hand being formed and present in the given crop to recover from tracking failure. If the score goes below a certain level, the detector is activated, and tracking is restarted. Another

important factor in excellent AR/VR hand interaction is handedness. This is especially useful in situations when each hand does its own set of tasks. As a consequence, a binary classification head is built to detect if the input hand is left or right. This system is geared for real-time mobile GPU inference, and lighter and heavier variations of the model were built to accommodate CPU inference on mobile devices without GPU capability, as well as greater accuracy requirements for desktop application. Figure 4 depicts a total of 20 distinct hand landmarks. Figure 5 depicts a variety of hand tracking photographs.

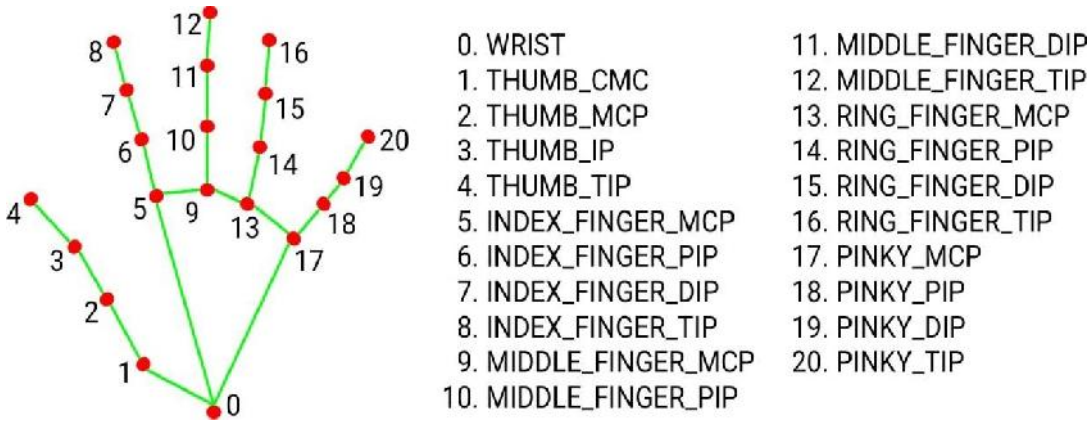

Figure 4. Hand Landmarks

Journal of Current Research in Engineering and Science
Bi-Annual Online Journal (ISSN : 2581 - 611X)
Volume 5- Issue 17, Paper 17 January 2022

Figure 5 Top: Aligned hand crops passed to the tracking network with ground truth annotation. Bottom: Rendered synthetic hand images with ground truth annotation.

5.3. Speech Recognition model

The Speech Recognition library is particularly adaptable since it functions as a wrapper for numerous major speech APIs. Microphone is used initially to gather data and then moved on to voice recognition. The Recognizer class is used to analyze the input command in this module. Various computer actions are performed with the aid of the processed data.

The Dialogue Module Manager functions similar to a chat bot, where the users write commands and the computer will execute different activities in response.

5. Implementation

The results obtained using various models are illustrated in this section. The Move Cursor gesture and the natural gesture are shown in Figure 6a and 6b. Move cursor can control the mouse. The user must spread his/her index finger and middle finger in V-shape to move the mouse. Natural Gesture is used to halt/stop execution of current gesture. When the user spreads all his five fingers the cursor does not move and it stops.

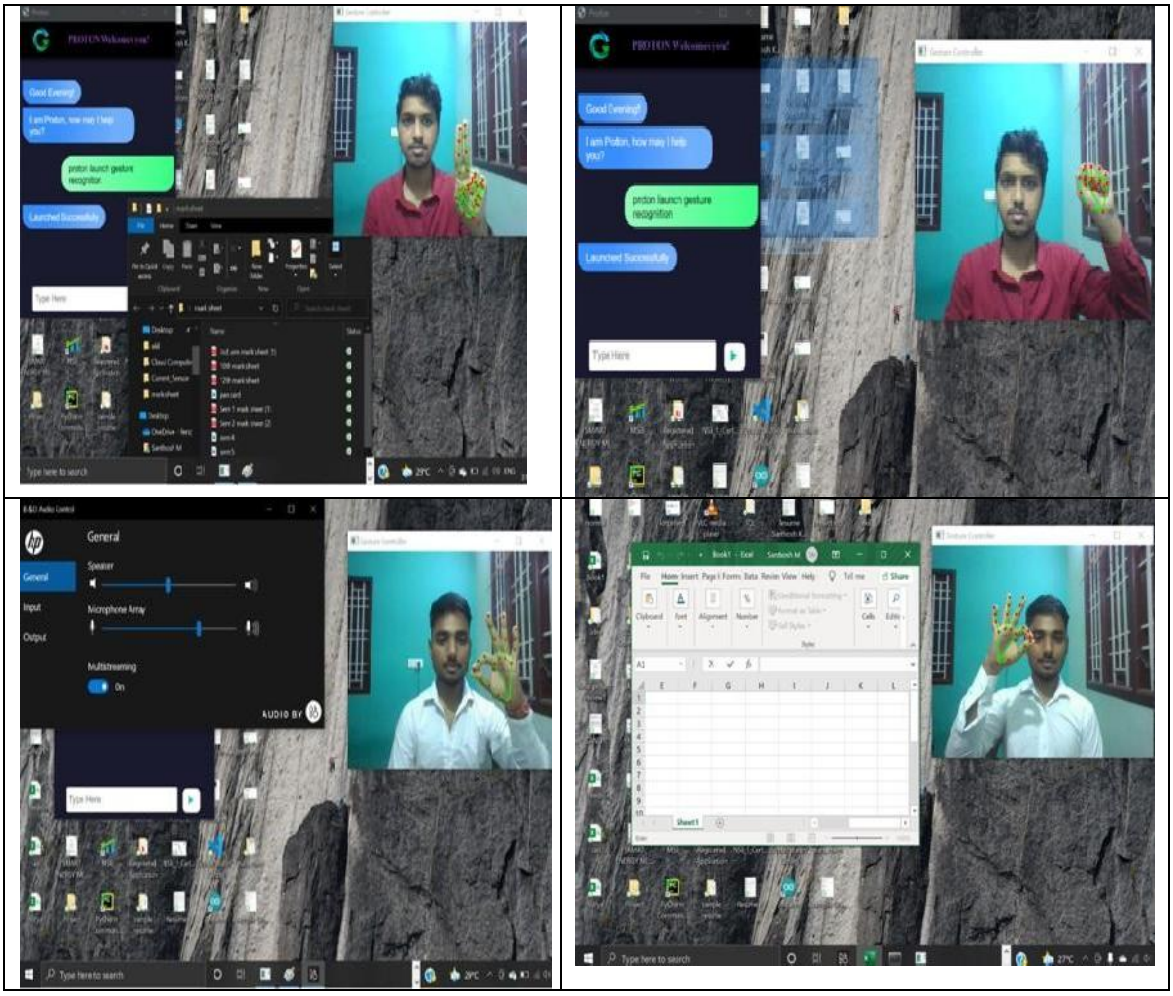

Figure 6 Move Cursor Gesture and Natural Gesture

When the user joins his/her index and middle finger the computer performs a double click depicted in Figure 6c. When the user closes all his fingers and moves across the screen the mouse is dragged and files are selected. This is the gesture for drag and drop and multiple object selection and is shown in Figure 6d. The same gesture can be used to move a file from one folder to another. If the user chooses a file and does this gesture the file is selected and when he releases his fingers in another folder the selected file is dropped there. The gesture for volume control is shown in figure 6e. The rate of increase/decrease of volume is proportional to the distance moved by pinch gesture from start point. When the gesture is moved up the volume increases and when it is moved down the volume decreases. Figure 6f shows the gesture for scrolling. This is the same gesture as shown in figure 6e but this gesture is done with left hand. The speed of scroll is proportional to the distance moved by pinch gesture from start point. Vertical and Horizontal scrolls are controlled by vertical and horizontal pinch movements respectively.

6. Conclusion

The above-mentioned proposed system analyses the effectiveness of technology in improving people's lives. There are various devices that aid in the restoration of better

lifestyles for people, but this project provides an improved version of existing systems. This project not only accepts commands by text, which is the most common mode of interaction, but also accepts input via voice commands and hand gestures in a more efficient and fluid manner than previous systems. It is designed to reliably recognize input and thus has a wide range of applications, ranging from photojournalism to medical technology to biometrics.

REFERENCES

[1] *Al-Hammadi, M., Muhammad, G., Abdul, W., Alsulaiman, M., Bencherif, M.A., Alrayes, T.S., Mathkour, H. and Mekhtiche, M.A, "Deep learning-based approach for sign language gesture recognition with efficient hand gesture representation", IEEE Access, 8,* ***(2020)*** *pp.192527-192542.*

[2] Alzahrani, M.S., Jarraya, S.K., Ben-Abdallah, H. and Ali, M.S.,"Comprehensive evaluation of skeleton features-based fall detection from Microsoft Kinect v2",*Signal, Image and Video Processing*,*13*(7), **(2019)** pp.1431-1439.

[3] Banerjee, A., Ghosh, A., Bharadwaj, K. and Saikia, H, "Mouse control using a web camera based on colour detection", **(2014)** *arXiv preprint arXiv:1403.4722*.

[4] Cui, J., Min, C. and Feng, D, "Research on pose estimation for stereo vision measurement system by an improved method: uncertainty weighted stereopsis pose solution method based on projection vector. *Optics express*, *28*(4), (**2020)**, p.5470- 5491.

[5] Duan, P., Wang, T., Cui, M., Sang, H. and Sun, Q., "Multi-person pose estimation based on a deep convolutional neural network" *Journal of Visual Communication and Image Representation*, *62*, **(2019)**, pp.245-252.

[6] Franco, A., Magnani, A. and Maio, D., "A multimodal approach for human activity recognition based on skeleton and RGB data". *Pattern Recognition Letters*, *131*, **(2020)**, pp.293-299.

[7] Jiang, D., Li, G., Sun, Y., Kong, J. and Tao, B., "Gesture recognition based on skeletonization algorithm and CNN with ASL database", *Multimedia Tools and Applications*, *78*(21), **(2019),** pp.29953-29970.

[8] Li, L., Chen, X., Wu, J., Wang, S. and Shi, G., "No-reference quality index of depth images based on statistics of edge profiles for view synthesis", *Information Sciences*, *516*, **(2020),** pp.205-219.

[9] Li, B., Liu, L., Shen, M., Sun, Y. and Lu, M., "Group-housed pig detection in video surveillance of overhead views using multi-feature template matching", *Biosystems Engineering*, *181*, **(2019),** pp.28-39.

[10] Nadeem, A., Jalal, A. and Kim, K., "Accurate physical activity recognition using multidimensional features and Markov model for smart health fitness", S*ymmetry*, *12*(11), **(2020),** p.1766.

[11] Plouffe, G. and Cretu, A.M., "Static and dynamic hand gesture recognition in depth data using dynamic time warping", *IEEE transactions on instrumentation and measurement*, *65*(2), **(2015),** pp.305-316.Xie, R. and Cao, J., "Accelerometer-based hand gesture recognition by neural network and similarity matching", *IEEE Sensors Journal*, *16*(11), **(2016)**, pp.4537- 4545.

[12] Zhang, Y. and Lu, X., "Measurement method for human body anteflexion angle based on image processing", *International Journal of Imaging Systems and Technology*,*29*(4), **(2019)**, pp.518-530.

[13] Zhang, W., Kong, D., Wang, S. and Wang, Z., "3D human pose estimation from range images with depth difference and geodesic distance", *Journal of Visual Communication and Image Representation*,*59*, **(2019),** pp.272-282.