

REAL-TIME HAND GESTURE DETECTION WITH 3-D KEY POINTS USING DEEP LEARNING

Sanjay R Pawar¹, Rameez Shamalik^{2*}, Nazim Mahammad Shaikh³, Pramod V Londhe⁴, Mohd Imran Ansari⁵, Ritesh G Mahajan⁶ and Sanjay Koli⁷

^{1,4}Department of Mechanical Engineering, Bharati Vidyapeeth (DU) College of Engineering, Pune

²Rameez Shamalik, Department of Computer Science and Engineering, School of Computing, MIT-ADT University, Pune, Maharashtra-412201, India

³Department of Basic Science and Humanities, Bharati Vidyapeeth (DU) College of Engineering, Pune

^{5,6}Department of Robotics & Automation Engineering, Bharati Vidyapeeth (DU) College of Engineering, Pune

⁷E&TC Dept., Ajeenkya D Y Patil school of engineering, Pune, India.

¹<https://orcid.org/0000-0002-4835-0347>, ²<https://orcid.org/0000-0002-3721-6011>, ³<https://orcid.org/0000-0001-6005-9387>

⁴<https://orcid.org/0000-0002-5151-818X>, ⁵<https://orcid.org/0000-0001-9477-6368>, ⁶<https://orcid.org/0000-0002-5911-987X>

⁷<https://orcid.org/0000-0001-5401-959X>

Email: srpawar@bvucoep.edu.in, *rameez.shamalik@mituniversity.edu.in, nmshaikh@bvucoep.edu.in, pvlondhe@bvucoep.edu.in, miansari@bvucoep.edu.in, rgmahajan@bvucoep.edu.in, sanjay.koli@dypic.in

ARTICLE INFO

Article History

Received: March 11, 2026

Reviewed: April 11, 2026

Accepted: May 11, 2026

Published: July 31, 2026

Keywords:

3D estimation,
Video processing,
Gesture detection,
Convolutional Neural Networks (CNN).

ABSTRACT

The importance of touchless systems for easier Human Machine Interaction (HMI) has been brought to light by recent pandemics. A key component of HMI, gesture detection represents a distinct class of useful computer vision applications. Outstanding outcomes in video processing are demonstrated by Convolutional Neural Networks (CNN) for a range of computer vision applications. This study reviews and compares various CNN versions for gesture detection along with their uses. To determine the correctness of the model, a local dataset of hand gestures is built and tested in conjunction with a synthetic gesture dataset. Using multilayered CNN, palm detection in a video frame is accomplished with efficient foreground-background separation. In a parametric research, these Mean Square Error (MSE) values are then contrasted with the known comparable system. The MSE of 12.6% obtained from the combination of synthetic and local datasets is better than that of recent research. In order to achieve 3 Dimensional (3D) estimation of a hand gesture with the aid of palm recognition, the number of key points allocated for the skeletal representation of the hand is crucial. This further enables the application of the technique in numerous upcoming genres.



Copyright ©2026 by authors and Galileo Institute of Technology and Education of the Amazon (ITEGAM). This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

I. INTRODUCTION

Convolutional Neural Networks (CNN's) are a subset of deep neural networks that are most frequently employed in image or video processing. One of the finest methods for extracting features from a video frame is convolutional networks. It recognises the supplied video frames as a pixel matrix, making it simpler to separate the contained items. Then, several levels with weights and biases are supplied with these pixels. These layers are all interconnected and aid in feature extraction before categorization. When gesture detection is only concerned with real-time video processing, a complex scenario is exhibited. The complexity of video processing algorithms has grown quickly over time [1], [2]. CNNs, because of their default categorization output, carry out the fundamental operation on videos to identify a certain gesture [3]. In accordance with the applications for which they are utilised, there are several types of neural networks. There are several different CNN designs, including as LeNet, AlexNet, VGG, GoogLeNet, ResNet, and many more. Since there are more variables in hand gesture detection than in face detection, it faces many challenges. Hands, unlike the face, which has fixed features such as the nose, eyes, and ears, have ten different shaped fingers. Their morphology and position fluctuate in real time, and gestures have a greater number of possible combinations.

There is a high likelihood that fingers will merge with the noisy background, as well as occlusion. Numerous researchers have been working on such confinements and developed 2D and 2.5D key point estimation techniques. This approach, while effective, lacks disparity and geometry for a real-time complex system. Semantic segmentation is essential for labelling pixels in a frame and aiding in the separation of the foreground and background. UNET and VGG-16, two CNN-based deep learning models, are created for segmentation and classification [4]. In order to assist astronauts in the space station with gesture detection as they deal with robots for their daily tasks, a combination of Mask-RCNN and MobileNet is employed for background removal and feature extraction, respectively [5]. Recently, the Hybrid Multidimensional Deep Convolutional Neural Network (HMDCNN) was created to enable multimodal recognition of speech as well as facial, lip, and even voice [6].

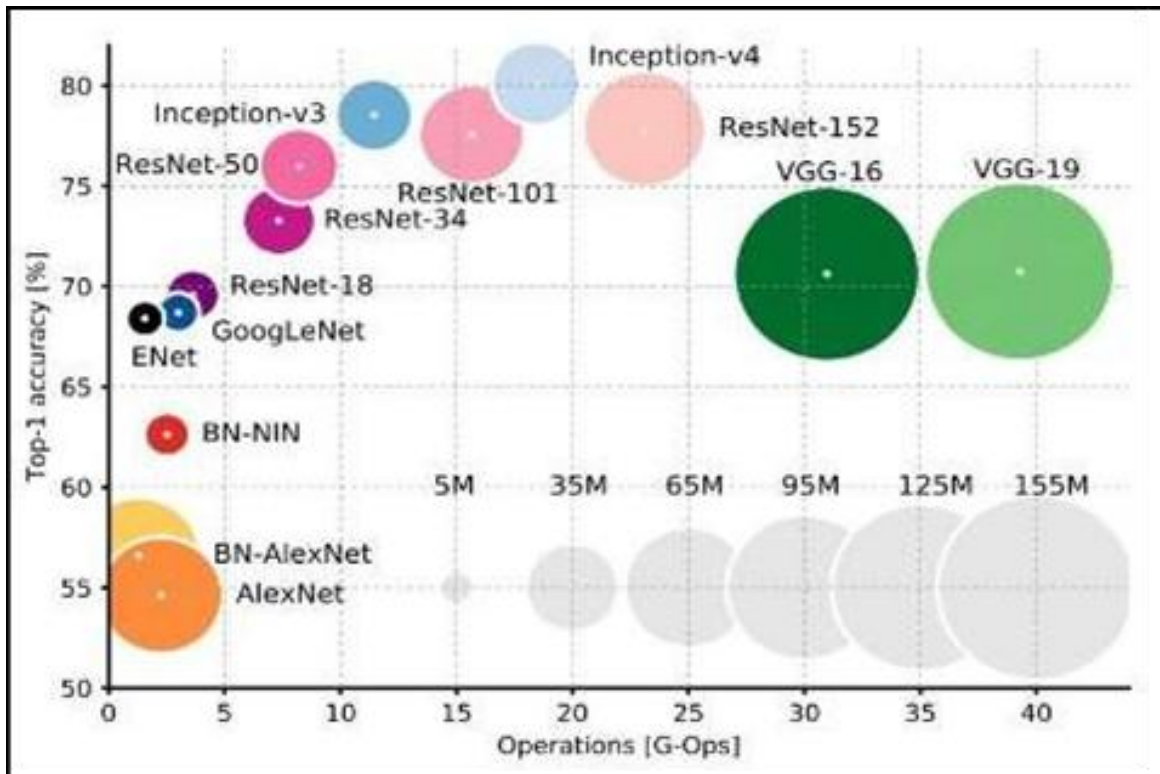


Figure 1: Accuracy Vs Number of operations performed in various types of CNNs.
Source: Authors, (2026).

Figure 1 illustrates how Alexnet and its variants provide respectable accuracy with regard to few operations carried out. The quantity of procedures is gradually raised as we go forward in our hunt for greater precision. When it comes to providing the best performance for a given amount of operations, Enet, GoogleNet, and ResNet are quite close. Processing video that comprises a lot of image frames quickly and efficiently is required for the main focus. However, even little flaws in video accuracy, such as lag or delay, utterly defeat the point. Newer versions of ResNet, Inception Net and Inverse Kinematics Net that provide a trade-off between accuracy and operations are helpful for solving such challenging real- world situations [7], [8]. Despite being rather hefty models, VGG-16 and VGG-19 provide good accuracy for video processing of about 70–75 percent while simultaneously processing millions of operations to evaluate the system's processing capabilities. For dynamic hand gesture identification, optical motion guided templates are used with 2D and 3D CNNs [9]. An easy-to-use but effective technique for extracting feature vectors for Chinese sign language identification uses 3D CNN [10] and optimised dense optical flow to analyse video stream.

When working with complicated datasets, it might sometimes be more productive to utilise a single network rather than the hybrid networks stated above. Therefore, a quicker R-CNN is employed with pose anchor key points for hand motions for the difficult Large scale Multiview 3D Hand Pose Dataset (LSM-HPD). Skeleton extraction from a video frame to identify motions and gestures is another innovative approach to this subject. To extract skeletal information and subsequently identify the action based on it, Recurrent Neural Networks (RNNs) and Long Short-Term Memory (LSTM) cells are utilised in conjunction [11]. The difficult challenge of recognising dynamic hand movements often requires the use of several CNN models. A 3D-Resnet network is employed for introduce fusion to abstract restricted spatiotemporal characteristics, and then a variation ConvLSTM is used to acquire global spatiotemporal data [12]. For the same application online, a CNN architecture with an effective sliding window technique is given [13]. It consists of a classifier and a detector on two levels. For efficiency analysis, a separate CNN is utilised. Figure 2 depicts the detector and classifier structure combination which uses single time activation function to save time and processing.

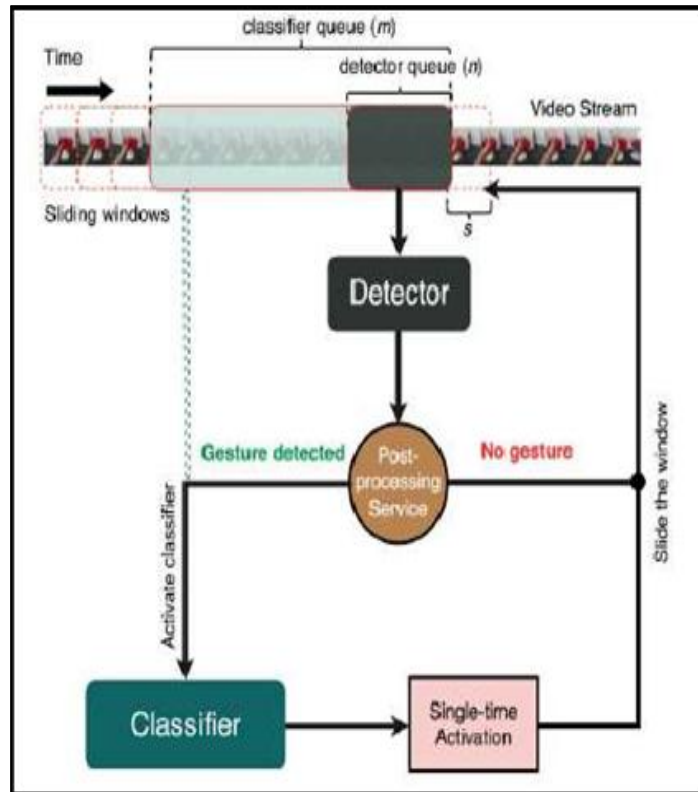


Figure 2: Two-level hierarchical structure.

Source: [13].

A glove sensor was introduced for Human Machine Interaction (HMI) in order to record the precise coordinates of hands. According to its intended use, gloves typically incorporate a variety of sensors, including an optical sensor [14], assistive technology for the blind and deaf [15], accelerometer and gyroscope sensors for human machine interfaces [16] or a color-based glove that can be customized with 3D models [17]. The main goal of computer vision is to do these tasks with the least amount of hardware possible by training a neural network, which adds more innovation and creativity to the desired job. In Methodology, it is demonstrated how to use a mixture of CNNs to obtain a 3D representation of gestures. In the gesture detection scheme, the palm represents the most stable stationary part of the hand. After acquiring video feed, the proposed research performs the Palm detection process first. The wrist of the palm is assumed to be the target value, and the key points are envisioned in relation to it. Lastly, joint estimation and prediction on a 3D surface are performed to replicate a 3D skeletal hand gesture instantaneously. The use of costly stereo sensors or multi camera set up at various perspectives is prevented, and a simple RGB camera is used instead.

In this paper the authors summarise their contributions as follows.

- a) An alternative model is suggested to resolve the problem
- b) Development of Local and Synthetic datasets
- c) Planned the model is trained and tested on these datasets
- d) Parametric study of final results with reference to latest contributions.

II. MATERIALS AND METHOD

With a few minor exceptions, all of the strategies described in the previous section utilise multilayer design. When real-time gesture detection is needed for video processing, assembling the necessary hardware is a laborious effort. When employing CNN for deep learning on such tasks, a graphics processing unit (GPU) is required. Herein study, real-time video processing is done on an NVIDIA GPU. Google Colab, Kaggle, and for paid or rented offerings Amazon Web Services (AWS) and Microsoft Azure enable access to GPU over the cloud, among many more operators, are some additional platforms that offer free GPU usage for a certain time frame. Therefore, the end output may be either a 3D dataset for model processing or a 3D estimate of gestures and actions in order to achieve greater accuracy [18]. A Google hand tracking resolution suggests an ML pipeline that combines two models that cooperate with one another: a palm detector and a hand landmark model [19]. Using an aligned hand bounding box, the former locates palms, and the latter yields high-fidelity 2.5D landmarks. Finding the depth from a particular frame is crucial to achieving 3D perception in computer vision. The easiest method of finding depth is to have users interact with the system [20]. Another approach may use the placement of numerous cameras [21] or possibly the use of a well-structured light to record the geometry of the frame in relation to a single image [22]. The most popular technique for doing this is monocular depth estimation, although for greater accuracy, more datasets are required [23]. A Gated Multi-Scale Network is employed to get the same result [24]. The time-consuming job of estimating depth, which calls for the use of numerous datasets, is substantially helped by the use of 3D key points.

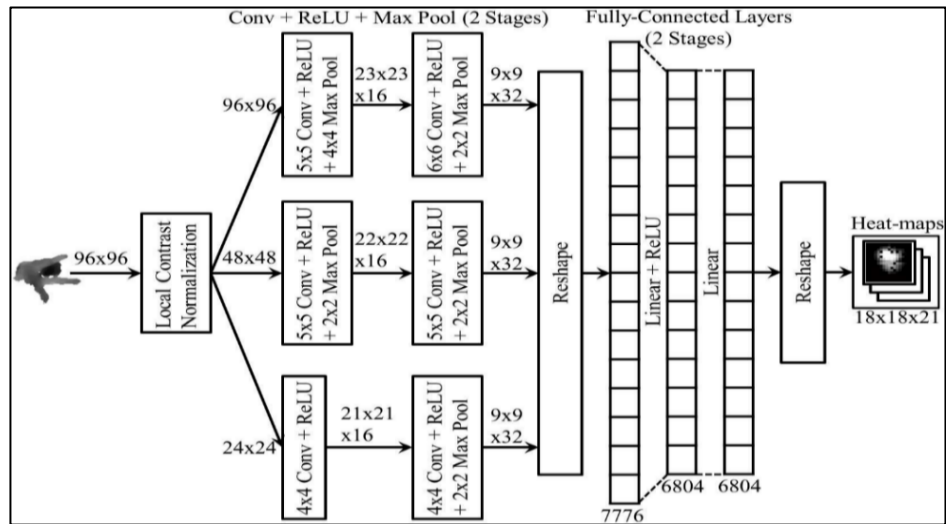


Figure 3: Combination of convolutional and fully connected layers.
Source: [25].

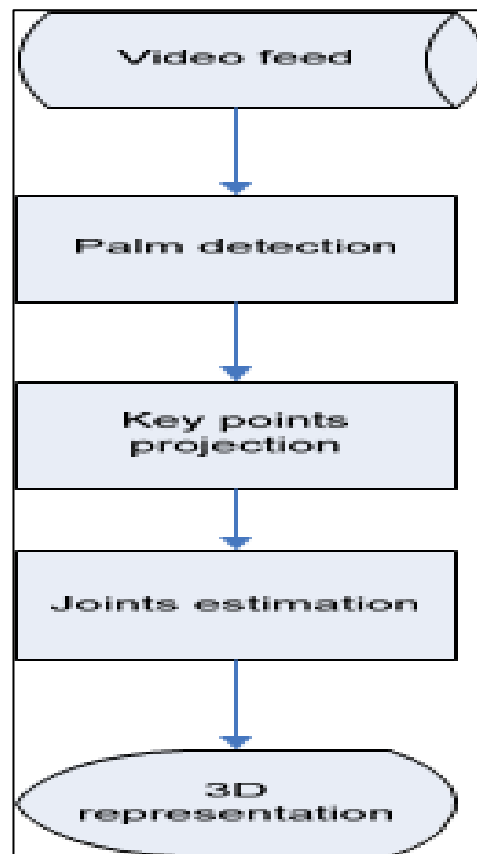


Figure 4: Flowchart of the proposed model.
Source: Authors, (2026).

As seen in figure 3, the gesture's primary point and focal length are determined after taking a depth reading from the relevant frame or picture, and the motion's local contrast is altered [25] which aids in the image normalisation process. The combination of 5 by 5 and 4 by 4 Convolution and ReLu is used along with 2 by 2 Max pooling. After being down sampled, the images are run through many convolutional layers. The feature map outputs are then input into fully linked layers to produce heat maps. As seen in figure 4, the only sensor employed in this suggested system is a camera, or video stream. The palm detector uses the video frames to form a convex hull around the hand, with the radius determining the borders of the fingers. The entire palm, including the fingers, is covered by the Euclidean distance which helps in Key points projection. The final phase projects the critical locations along the defined contour. In order to connect the important locations in a skeleton-like form, the Gaussian distribution of joints for Joints estimation is determined using three views of heat maps. Equation 1 may be used to determine a 3D representation of joint positions at the final stage.

II.1 STEPWISE IMPLEMENTATION

Where , a measure of reconstruction of is optimal joints, suggests optimal coefficients for joints, is total amount of principal components included and mean vector constant. It has been observed that 21 key points, as opposed to other number of key points, provide better throughput and are linked to each other for precise gesture and geometry estimation.

II.2 DATASETS AND PRE-PROCESSING

Two distinct dataset kinds are joined to get the desired outcome. Starting with 5k images made from locally gathered gesture dataset [26], each with a unique collection of gestures. Each image has a different backdrop, which adds to the diversity of frames available for training the model. Aside from the front motions that are necessary, the background has intricate shape and variable illumination. Additionally, a synthetic dataset with more than 50k images that includes hand movements with various skin tones is employed. To accurately define the skeleton hand shape, annotations are pre-processed on a local gesture dataset with a maximum of 21 key points. These crucial details makes it easier to capture the widest variety of hand gestures and movements possible in any one frame. With the aid of 21 crucial points, the skeleton hand may be used to identify motions, without the need for bright illumination or an unsettling background.

II.3 TRAINING DETAILS

The hyper parameters are chosen in such a way that it reaches a trade-off between the complexity of the model and its likely results. The two models are trained using the Root Mean Square Propagation (RMSProp) optimizer, which has a learning rate of 10-4. Both modules have 50 iterations, with the first module having a batch size of 32 and the second having a batch size of 64. The model is trained using 75% of the photos, tested using 15%, and then validated using the remaining images. The system is trained considering multiple losses which includes Regularisation loss (E_{reg}), 3D loss (E_{3D}), Normalisation loss (E_{Norm}) and Angular loss (E_{Ang}) as shown in equation 2.

$$E_{total} = E_{reg} + E_{3D} + E_{Norm} + E_{Ang} \quad (2)$$

III. RESULTS AND DISCUSSIONS

Figures 5 (a) and (b) demonstrate how our model, even in low light and noisy background, can recognise the palm in the supplied video frame while removing the complicated backdrop. Even though it cannot precisely trace the critical locations on the index finger in 5 (a), the 3D graph nevertheless provides a better depiction. Joint estimation is crucial in this case to project the skeleton form of the palm and fingers, and the gesture's 3D structure amplifies the geometry and depth of the gesture in the available space. The reconstruction of hands in a three-dimensional axis improves the accuracy and precision of gestures in real-time. To analyse the gesture in a particular video frame from various perspectives, the hands XYZ axis representation may be rotated.

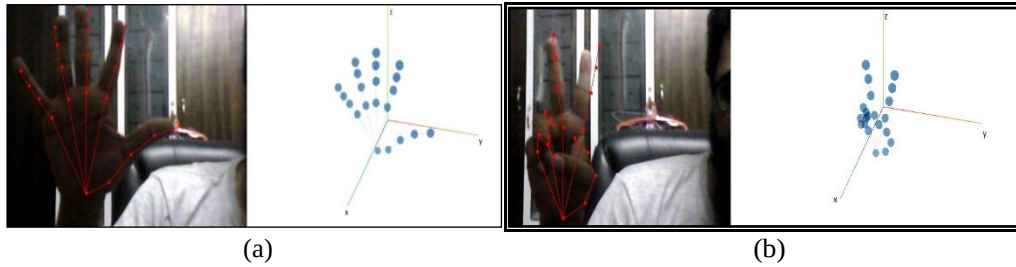


Figure 5: (a) & (b) Shows key points projected on palm, in low light and noisy background with its skeletal representation in XYZ axis. Source: Authors, (2026).

The simulated dataset demonstrates improved accuracy for locating the precise critical locations in palm. Since the model must function in real-world situations, the locally produced hand gesture dataset is more important since artificial motions, in contrast to real-world situations, have smooth skin textures and steady backdrops. It is therefore necessary to combine the two datasets, which improves accuracy in terms of Mean Square Error (MSE), which provides a measure of the quality of frame predictions, when compared to utilising only the local dataset with authentic hand gestures. Mean square error represents the difference of the squared error between actual (A_v) and predicted (P_v) output as shown in equation 3.

$$MSE = \frac{1}{n} \sum_{i=1}^n (A_v - P_v)^2 \quad (3)$$

To analyse the gesture in a particular video frame from various perspectives, the hands XYZ axis representation may be rotated. A trade-off between aims and relevance is developed as a result. MSE should ideally be zero and practically as low as possible. The suggested technique outperforms the corresponding traditional model, as demonstrated in table 1, when compared. The Local Hands dataset gives MSE of 15.7 while synthetic Hands gives 25. On the other hand combination of both datasets gives superior MSE of 12.6, all in terms of percentage.

Table 1: Model performance for available datasets.

Datasets	Proposed System MSE (%)	FanZengh et.al., MSE (%)
Local Hands	15.7	16.1
Synthetic Hands	25	25.7
Combination of Both	12.6	13.4

Source: Authors, (2026).

Percentage of Correct Key-points (PCK) is another method to measure the amount of the hand gesture key points and joint estimation. When the difference between the predicted and the true joint is smaller than some threshold, the identified joint is considered to be correct.

To come up with the distances, the pixel lengths in each case are normalised by 0.7 the corresponding hand size of the corresponding person, with regard to threshold which are applied in the development of the PCK curves. The blue and green dotted lines are the PCK of local and synthetic hand gesture dataset respectively and the red line is the PCK of mixture of both datasets. In this way, the tradeoff can be well brought out in terms of practical use of hand gesture datasets.

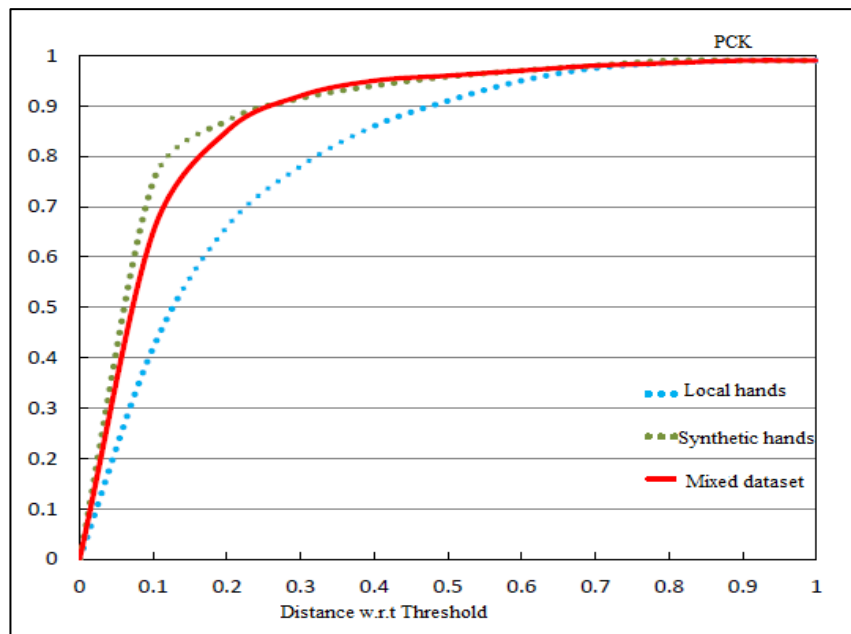


Figure 6: PCK comparison between different datasets.

Source: Authors, (2026).

IV. CONCLUSIONS

Real-time gesture recognition is well-suited for CNN operations, particularly in feature extraction and classification.

- Various CNN configurations can be employed for dynamic, online gesture detection.
- Combining results from two separate CNNs tends to outperform a single CNN in gesture recognition.
- The use of complex hardware enhances computer vision solutions for real-time video processing.
- Locally created dataset exhibits moderate accuracy with 15.7% MSE with actual video frames.
- Research utilizes a combination of datasets for training and testing, resulting in improved accuracy and a reduced MSE of 12.6%.
- PCK is a parameter used for exact projection of key points around hand gestures, showing consistent results on both datasets.
- Gesture sensing technologies find applications in gaming, virtual reality, and augmented reality.

V. AUTHOR'S CONTRIBUTION

Conceptualization: Sanjay R Pawar and Rameez Shamalik.

Methodology: Sanjay R Pawar and Mohd Imran Ansari.

Investigation: Sanjay R Pawar and Nazim Mohammad Shaikh.

Discussion of results: Sanjay R Pawar, Rameez Shamalik, Ritesh G Mahajan and Mohd Imran Ansari.

Writing – Original Draft: Sanjay R Pawar.

Writing – Review and Editing: Sanjay R Pawar, Rameez Shamalik and Ritesh G Mahajan.

Resources: Sanjay R Pawar and Ritesh G Mahajan.

Supervision: Mohd Imran Ansari and Rameez Shamalik.

Approval of the final text: Sanjay R Pawar, Pramod V Londhe, Sanjay Koli and Rameez Shamalik.

VI. REFERENCES

- [1] Wiedemann, S., Kirchhoffer, H., Matlage, S., Haase, P., Marbán, A., Marinc, T., Neumann, D., Nguyen, T., Schwarz, H., Wiegand, T., Marpe, D., & Samek, W. (2019). DeepCABAC: A Universal Compression Algorithm for Deep Neural Networks. *IEEE Journal of Selected Topics in Signal Processing*, 14, 700-714.
- [2] Koli, S.M., & Shamalik, R. (2019). Transformation of Video Signal Processing Techniques from 2D to 3D: A Survey. *Lecture Notes in Electrical Engineering*. 570, pp.63–70.
- [3] Shamalik, R., & Koli, S.M. (2022). Real Time Gesture Detection Using Convolutional Neural Network. 2022 IEEE 4th International Conference on Cybernetics, Cognition and Machine Learning Applications (ICCCMLA), 1-4.
- [4] Dutta, H.P., Sarma, D., Bhuyan, M.K., & Laskar, R.H. (2020). Semantic Segmentation based Hand Gesture Recognition using Deep Neural Networks. 2020 National Conference on Communications (NCC), 1-6.
- [5] Lingyun, G., Lin, Z., & Zhaokui, W. (2020). Hierarchical Attention-Based Astronaut Gesture Recognition: A Dataset and CNN Model. *IEEE Access*, 8, 68787-68798.
- [6] Vynokurova, O., & Peleshko, D. (2020). Hybrid Multidimensional Deep Convolutional Neural Network for Multimodal Fusion. 2020 IEEE Third International Conference on Data Stream Mining & Processing (DSMP), 131-135.

- [7] Shamalik, R., & Koli, S.M. (2023). Effective and efficient approach for gesture detection in video through monocular RGB frames. *Multimedia Tools and Applications*, 82, 17231-17242.
- [8] Shamalik, R., & Koli, S. (2022). DeepHands: Dynamic hand gesture detection with depth estimation and 3D reconstruction from monocular RGB data. *Sādhana*, 47, 1-9.
- [9] Sarma, D., Kavyasree, V., & Bhuyan, M.K. (2022). Two-stream fusion model using 3D-CNN and 2D-CNN via video-frames and optical flow motion templates for hand gesture recognition. *Innovations in Systems and Software Engineering*, 1 - 14.
- [10] Zhao, K., Zhang, K., Zhai, Y., Wang, D., & Su, J. (2020). Real-Time Sign Language Recognition Based on Video Stream. 2020 39th Chinese Control Conference (CCC), 7469-7474.
- [11] Avola, D., Cascio, M., Cinque, L., Foresti, G.L., Massaroni, C., & Rodolà, E. (2020). 2-D Skeleton-Based Action Recognition via Two-Branch Stacked LSTM-RNNs. *IEEE Transactions on Multimedia*, 22, 2481-2496.
- [12] Peng, Y., Tao, H., Li, W., Yuan, H., & Li, T. (2020). Dynamic gesture recognition based on feature fusion network and variant ConvLSTM. *IET Image Process.*, 14, 2480-2486.
- [13] Köpüklü, O., Gunduz, A., Kose, N., & Rigoll, G. (2020). Online Dynamic Hand Gesture Recognition Including Efficiency Analysis. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2, 85-97.
- [14] Fujiwara, E., Ferreira Marques Dos Santos, M., & Suzuki, C.K. (2014). Flexible Optical Fiber Bending Transducer for Application in Glove-Based Sensors. *IEEE Sensors Journal*, 14, 3631-3636.
- [15] Shrote, S.B., Deshpande, M., Deshmukh, P., & Mathapati, S. (2014). Assistive Translator for Deaf & Dumb People.
- [16] Gupta, H.P., Chudgar, H.S., Mukherjee, S., Dutta, T., & Sharma, K. (2016). A Continuous Hand Gestures Recognition Technique for Human-Machine Interaction Using Accelerometer and Gyroscope Sensors. *IEEE Sensors Journal*, 16, 6425-6432.
- [17] Lamberti, L., & Camastra, F. (2011). Real-Time Hand Gesture Recognition Using a Color Glove. *International Conference on Image Analysis and Processing*.
- [18] Shamalik R, Koli S, (2020). Emergence and Functionality of 3D Videos. *International Journal of Engineering and Advanced Technology*.
- [19] Zhang, F., Bazarevsky, V., Vakunov, A., Tkachenka, A., Sung, G., Chang, C., & Grundmann, M. (2020). MediaPipe Hands: On-device Real-time Hand Tracking. *ArXiv*, abs/2006.10214.
- [20] López, A., Garcés, E., & Gutierrez, D. (2014). Depth from a Single Image Through User Interaction. *Spanish Computer Graphics Conference*.
- [21] Shamalik, R., & Koli, S.M. (2021). Real Time Human Gesture Recognition: Methods, Datasets and Strategies. *Recent Trends in Intensive Computing*.
- [22] Bahmutov, G., Popescu, V., & Sacks, E. (2004). Depth Enhanced Panoramas. *IEEE Visualization 2004*, 11p-11p.
- [23] Ranftl, R., Lasinger, K., Hafner, D., Schindler, K., & Koltun, V. (2019). Towards Robust Monocular Depth Estimation: Mixing Datasets for Zero-Shot Cross-Dataset Transfer. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44, 1623-1637.
- [24] Lin, L., Huang, G., Chen, Y., Zhang, L., & He, B. (2020). Efficient and High-Quality Monocular Depth Estimation via Gated Multi-Scale Network. *IEEE Access*, 8, 7709-7718.
- [25] Ge, L., Liang, H., Yuan, J., & Thalmann, D. (2016). Robust 3D Hand Pose Estimation in Single Depth Images: From Single-View CNN to Multi-View CNNs. 2016 *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 3593-3601.
- [26] Shamalik, R., & Koli, S. (2023). "FabDepth I: A Unique Dataset for Efficient Gesture Detection". *International Journal of Information Technology*, 15, 2645 - 2649.