Convolution Neural Network based Hand Gesture Recognition System

Shalvee Meshram
Dept. of ESE, National Institute of Electronics, and Information Technology Aurangabad, Maharashtra, India
shalveepm@gmail.com

Roshan Singh
Dept. of ESE, National Institute of Electronics, and Information Technology Aurangabad, Maharashtra, India
rsrajput1245@gmail.com

Prashant Pal
Scientist B
National Institute of Electronics and Information Technology
Aurangabad, Maharashtra, India
prashantpal@nielit.gov.in

Shashank Kumar Singh
Scientist B
National Institute of Electronics and Information Technology
Aurangabad, Maharashtra
shashank@nielit.gov.in

2023 Third International Conference on Advances in Electrical, Computing, Communication and Sustainable Technologies (ICAECT) | 978-1-6654-9400-7/23/$31.00 ©2023 IEEE | DOI: 10.1109/ICAECT57570.2023.10118267

***Abstract*— Strong hand gesture recognition has been essential in the area of human-computer interaction for a very long time. Due of their intricacy and breadth, these gestures are difficult for many people to understand, which hinders communication between those with and without speech impairments. There is a lot of active practical research being done in the field of computer vision because of the current boom in deep learning. Till the date, various image processing algorithms have used color and depth cameras to recognize hand gestures, however it is still difficult to classify movements from different subjects accurately. The objective of this study is to identify hand gestures by using a camera to quickly follow the region of interest (ROI), which in this case is the hand region, in the image range. In this project, the use of convolutional neural networks (CNNs) in an algorithm for real-time hand gesture recognition has been proposed in this paper. On a dataset of thirty-six hand gestures and 400 photos for each gesture, the suggested CNN is anticipated to achieve excellent accuracy.**

Keywords—convolutional neural network, sign language, hand gesture recognition, deep learning, machine learning

I. Introduction

Artificial intelligence and robotics have been used in recent years to augment the autonomy of people with disabilities. The major goal in this situation is to enhance quality of life by enabling users to complete a larger variety of daily duties more effectively. Particularly for Sign Language Recognition (SLR), hand gesture recognition has been acknowledged as a beneficial technology for several application domains [2], [3], [4], [5], [6]. Complex hand gestures are used in sign languages, and even tiny hand movements can convey a wide range of meanings. This led to the introduction of numerous vision-based dynamic hand gesture detection systems over the past ten years. Certain researchers used, the Hand Gesture Translation System (HGT System) [16], [17] that enables the deaf and dumb to interact with ones that can hear and speak, using a variety of hand gestures. The deaf and dumb will be able to communicate with people much more swiftly by using these hand signals. The output is created in text format so that the individuals can communicate with other people.

Recent recognition tasks have been successfully classified using deep convolutional neural networks. It has been demonstrated that multi-column deep CNNs, which use several parallel networks, can increase recognition rates of single networks by up to 80% for a variety of image classification tasks. Karpathy et. a1 [10] found that combining CNNs trained with two different streams of the original and spatially cropped video frames produced the best results for large-scale video categorization.

In this project, the goal is to maintain system accuracy and speed while recognizing a collection of static and dynamic hand motions. The computer is controlled by recognized gestures. A CNN-based classifier is constructed for hand shape recognition by applying transfer learning to a convolutional neural network that has already been trained on a large dataset. A method that uses 2D convolutional neural networks to learn and predict while extracting hand components from the image has been developed. An efficient spatio-temporal data augmentation strategy to distort the input volumes of hand gestures is thus suggested to decrease potential over-fitting and enhance generalization of the gesture classifier. Existing spatial augmentation methods are also included in the augmentation method.

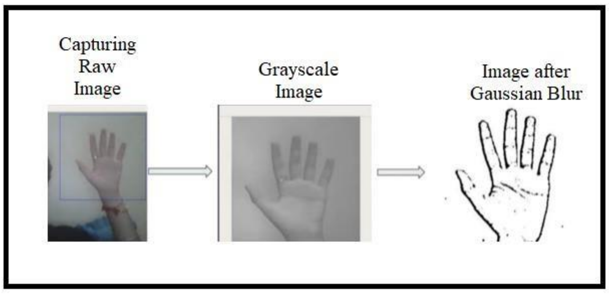


Fig. 1

II. Related Work

There has been a lot of research done on hand gesture recognition up to this point, but there has not been any good research done on sign language alphabets. A database of 240 photos for each of the twenty-four sign language symbols was produced by Singha et al. From the photos, they retrieved eigen values, which they then categorized based on Euclidean distance. Their program was able to classify photos with a 97% accuracy, however the images could only have a limited number of static backdrop circumstances and gestures. Liao et al., on the other hand, used an Intel Real Sense RGB-Depth sensor in his project and used depth perception methods to separate the hand region from the background.

III. METHODOLOGY

A. Workflow

Transfer learning is used here to train a CNN-based classifier for hand shape identification over a pretrained convolutional neural net that was previously trained on a sizable dataset. The pretrained model that we are using is VGG16.
Each frame is fed into the classifier after resizing and padding. The commanding phase begins right away if the classified hand makes a static motion. If not, the process of hand tracing is next. Fig. 2 displays the block diagram of our suggested technique.

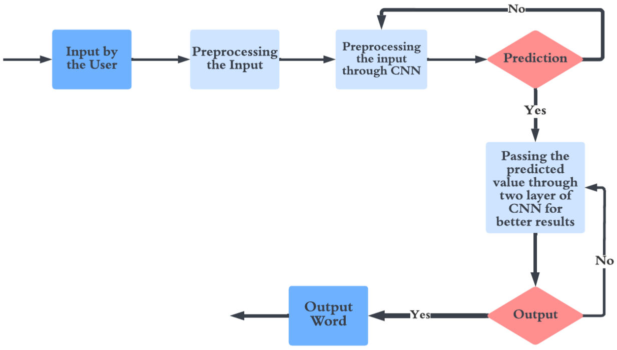

Input by the User
Preprocessing the Input
Preprocessing the input through CNN
Prediction
No
Yes
Passing the predicted value through two layer of CNN for better results
No
Output
Yes
Output Word

Fig. 2

B. Dataset

A training set is required for model training. For the aim of training a model, gestures are created manually through python programming via webcam. The hand shape is all that is needed to recognize a static gesture.
The trained classifier issues an instruction to the computer after classifying the hand form as a static gesture. Dynamic gestures, in contrast to static gestures, call for both hand motion and shape. The hand area is segmented off using the HSV (Hue, Saturation, Value) skin color algorithm in a frame for tracing dynamic hand gestures, and then the blob area is cropped. The blob's centroid is located and tracked. The major goal of this step is to locate the traced hand's center in each frame by collecting its coordinates. These coordinates will be utilized to determine which computer command and which motion go together.

In this project, a dataset of 400 images of 36 hand gestures (27 new hand gestures and 9 older hand gestures that were used in the existing system) has been acquired using a webcam to evaluate the model. Each image is a 50x50 pixels. Black and white skin pixels are created by removing color skin pixels from the color image by the method called Gaussian Blur (Fig.1 and Fig. 3). In a Gaussian Blur operation, a Gaussian filter [29] is fused with the image rather than a box filter.

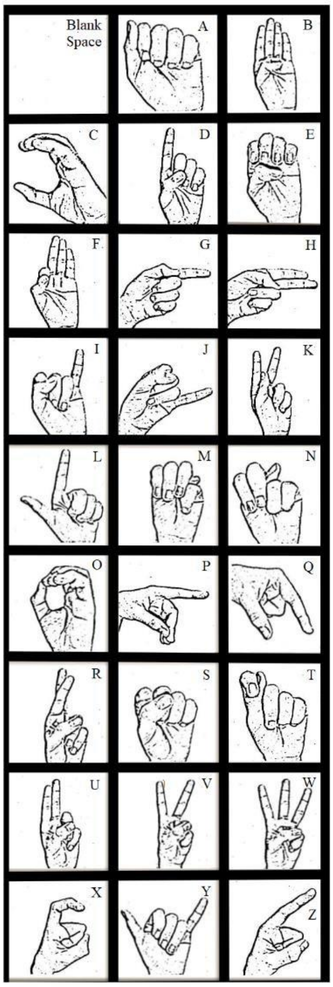


Fig. 3 Dataset of Hand Gestures of alphabets A-Z in American Sign Language

The high-frequency components are removed using the low-pass Gaussian filter. These black-and-white pictures are later cropped to 50x50 pixels.

Additional elements, such as empty spaces and background objects, are present in the frame that contains a gesture. For a better outcome, these extra, unused pieces must be separated. To remove the hand gesture, the obtained part is trimmed. Three stages are taken to complete the cropping. First, global thresholding is used to transform the cropped frame to binary, which is black and white. The second step is to remove the object of interest from the frame.
Each hand gesture's associated images are organized into a different folder. Each folder has a text file with entries for each of the images inside. One of the hand gestures seen in the image (fig.3) is indicated by an entry in the text file. 4000 extra photos have been obtained using spatio-temporal data augmentation techniques in addition to this dataset. The sections that follow explore the technique in more detail.

C. Classifier

The primary objective of convolution is to extract features from the input, such as edges, colours, and corners. As we delve further into the network, it begins to recognise more complex elements like shapes, numbers, and even individual facial features.

Six 2D convolution layers make up the network, and a max-pooling operator comes after each layer. The volumes at each layer, the convolution kernel sizes, and the pooling operators are all displayed in Fig. 4.

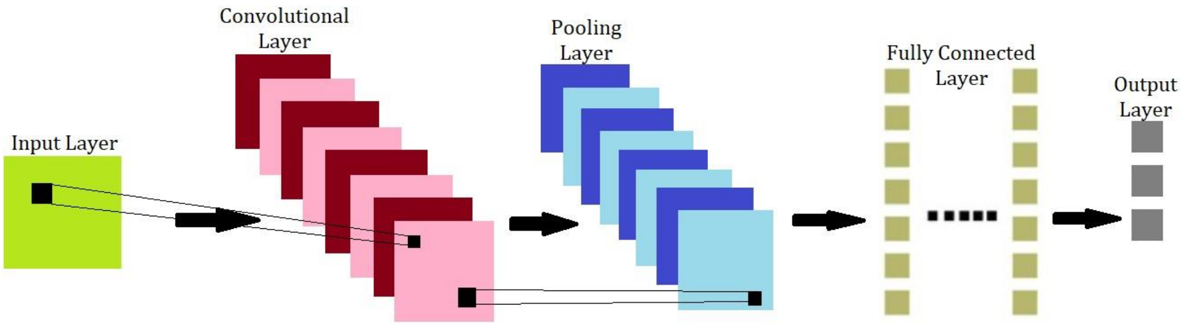


Fig.4 2D CNN Architecture

A fully linked network with 36 layers receives the output of the sixth convolution layer as input. Apart from the last output layer, which includes 36 neurons—one for each of the 36 hand gestures—each layer has 512 hidden neurons.

In the output layer, a sigmoid (logistic) activation function [1], [28] is applied. The sigmoid function is calculated as an "S" graph, wherein the input value for x ranges from 0 to 1. This is how the sigmoid activation function is calculated:

$$\boxed{A = 1/1+e^{-x}}$$

The remaining 35 levels employ the Tanh (hyperbolic tangent) activation function, which is a sigmoid function that has been mathematically adjusted. The input values for the function range from -1 to +1 and the function is represented as:

$$\boxed{\mathrm{Tanh} = e^{x} - e^{-x} / e^{x}}$$

The acquisition of a big dataset for each subject in the context of this article would be time-consuming and impractical when considering real-life applications, as a user would frequently not put up with hours of data recording for each training. In the subsections that follow, Batch Normalization is used to address this overfitting problem in more depth.

C.1 Batch Normalization

An algorithmic technique called batch-normalization (BN) [13] speeds up and improves the stability of Convolutional Neural Networks (CNN) training. BN normalises each batch of data through each layer during training. Once the network has finished training, the data is sent through one final time to compute the data statistics in a layer-by-layer method that are then set at test time. It was demonstrated that BN produced quicker training times while enhancing system correctness and regularisation. The first and second statistical moments (mean and variance) of the current batch are used to normalise activation vectors from hidden layers [24]. This normalisation step, suggestively, is used just prior to (or immediately following) the nonlinear function.

D. Training

To reduce a cost function for the dataset, network parameters must be optimized as part of the training process for a CNN. This project used the root mean squared method for the same. Liao et al. simultaneously trained segmented depth and RGB pictures using a double channel convolutional neural network-based architecture. The suggested model used a similar approach to simultaneously train the RGB-D pairs. We separately trained RGB and depth pictures using the same CNN-based architecture depicted in Fig. 4 and evaluated its performance both offline and in real time.

Further optimization was carried out using stochastic gradient descent. At each iteration, the Nesterov accelerated gradient was used to update the network's parameters. A set of random samples was used to initialize the weights of the 2D convolutional layers. The following subsections go into further information about these terms.

If the cost function did not improve by more than 10% in the previous 40 epochs, we reduced the learning rate by a factor of 2 and reset the rate to 0:005. After the learning rate had decreased by at least 4 times or if there had been more than 300 epochs, network training was then stopped. The network configuration with the minimum error on the training set was further chosen.

D1. Stochastic Gradient Descent

Iteratively minimizing an objective function that is expressed as the sum of differentiable functions, stochastic gradient descent (SGD), sometimes referred to as incremental gradient descent, is a stochastic approximation of gradient descent optimization. To put it another way, SGD iteratively searches for minima or maxima.

E. Spatio- Temporal Data Augmentation

The act of identifying intriguing, previously unidentified, but potentially helpful patterns from data gathered over time and location is known as spatial and spatiotemporal data mining. In this project, the dataset for training only contains 14,400 gestures, which is insufficient to avoid overfitting. Spatial and temporal data augmentation is done to prevent overfitting.

IV. Results

In this project, a live video camera records hand gestures. However, this presents a challenge when using the same software in various lighting scenarios. The RGB (Red, Green, Blue) image is transformed into an HSV (Hue, Saturation, Value) image for hand gesture identification. The next step is thresholding, which lowers background noise.

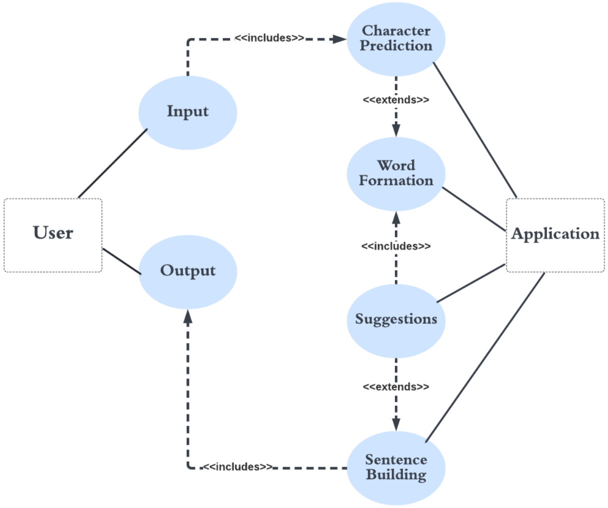
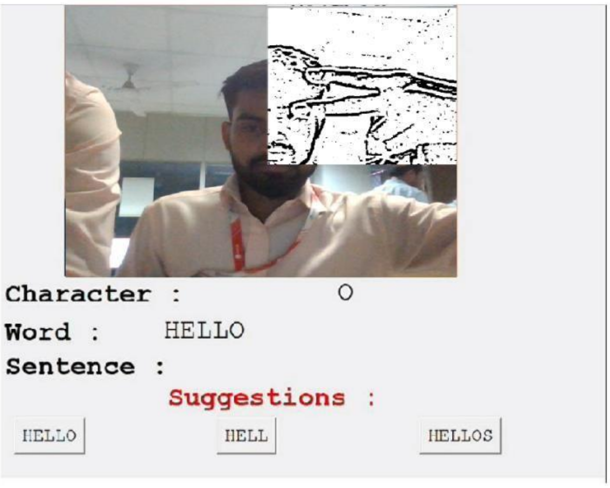


Fig. 5

	Dataset	Trained Data: Tested Data	CNN Model	Accuracy
Existing System [30]	4500	7:3	2D CNN	98.74%
Proposed System	14,400	6:4	2D CNN + VGG16	99.08%

However, if the lighting in the room changes, we must manually adjust the HSV value to obtain a satisfactory result, which complicates activities. Due to differences in skin tone, hand gestures from various hands can occasionally be problematic.

After randomising the data collected from all five subject areas, the dataset was split 70:30 between the training set and the testing set. To train static-based gestures, artificial data synthesis was incorporated in real time. Because artificial data synthesis was not used for the testing set, which is significantly easier to predict than the training set, the testing accuracy is higher than the training accuracy.



Character : O

Word : HELLO

Sentence :

Suggestions :

V. Conclusion

With the help of 2D convolutional neural networks, an efficient approach for recognising dynamic hand gestures has been created. To prevent overfitting, the suggested classifier augments the data with spatiotemporal information. It has been proven through comprehensive examination that combining low and high resolution sub-networks significantly increases classification accuracy. Additionally, it is shown that the suggested data augmentation strategy is crucial for getting improved performance. Thus, the system we created is comparatively noble to the existing systems with a higher accuracy and added elements that will eliminate issues of overfitting as well as disturbing light scenarios.

References

[1] N. A. Ahmad, "A Globally Convergent Stochastic Pairwise Conjugate Gradient-Based Algorithm for Adaptive Filtering," in IEEE Signal Processing Letters, vol. 15, pp. 914-917, 2008, doi: 10.1109/LSP.2008.2005437.

[2] S. Mitra and T. Acharya, "Gesture Recognition: A Survey," in IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 37, no. 3, pp. 311-324, May 2007, doi: 10.1109/TSMCC.2007.893280.

[3] N. H. Dardas and N. D. Georganas, "Real-Time Hand Gesture Detection and Recognition Using Bag-of-Features and Support Vector Machine Techniques," in IEEE Transactions on Instrumentation and Measurement, vol. 60, no. 11, pp. 3592-3607, Nov. 2011, doi: 10.1109/TIM.2011.2161140.

[4] P. Molchanov, X. Yang, S. Gupta, K. Kim, S. Tyree and J. Kautz, "Online Detection and Classification of Dynamic Hand Gestures with Recurrent 3D Convolutional Neural Networks," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 4207-4215, doi: 10.1109/CVPR.2016.456.

[5] P. Molchanov, S. Gupta, K. Kim and J. Kautz, "Hand gesture recognition with 3D convolutional neural networks," 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), 2015, pp. 1-7, doi: 10.1109/CVPRW.2015.7301342.

[6] Ming-Hsuan Yang, N. Ahuja and M. Tabb, "Extraction of 2D motion trajectories and its application to hand gesture recognition," in IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 8, pp. 1061-1074, Aug. 2002, doi: 10.1109/TPAMI.2002.1023803.

[7] A. S. Ghotkar, R. Khatal, S. Khupase, S. Asati and M. Hadap, "Hand gesture recognition for Indian Sign Language," 2012 International Conference on Computer Communication and Informatics, 2012, pp. 1-4, doi: 10.1109/ICCCI.2012.6158807.

[8] S. K. Shareef, I. V. S. L. Haritha, Y. L. Prasanna and G. K. Kumar, "Deep Learning Based Hand Gesture Translation System," 2021 5th International Conference on Trends in Electronics and Informatics (ICOEI), 2021, pp. 1531-1534, doi: 10.1109/ICOEI51242.2021.9452947.

[9] S. Hussain, R. Saxena, X. Han, J. A. Khan and H. Shin, "Hand gesture recognition using deep learning," 2017 International SoC Design Conference (ISOCC), 2017, pp. 48-49, doi: 10.1109/ISOCC.2017.8368821.

[10] A. Karpathy, G. Toderici, S. Shetty, T. Leung, R. Sukthankar and L. Fei-Fei, "Large-Scale Video Classification with Convolutional Neural Networks," 2014 IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1725-1732, doi: 10.1109/CVPR.2014.223.

[11] Jie Huang, Wengang Zhou, Houqiang Li and Weiping Li, "Sign Language Recognition using 3D convolutional neural networks," 2015 IEEE International Conference on Multimedia and Expo (ICME), 2015, pp. 1-6, doi: 10.1109/ICME.2015.7177428.

[12] V. de Oliveira Silva, F. de Barros Vidal and A. R. Soares Romariz, "Human Action Recognition Based on a Two-stream Convolutional Network Classifier," 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA), 2017, pp. 774-778, doi: 10.1109/ICMLA.2017.00-64.

[13] V. Thakkar, S. Tewary and C. Chakraborty, "Batch Normalization in Convolutional Neural Networks — A comparative study with CIFAR-10 data," 2018 Fifth International Conference on Emerging Applications of Information Technology (EAIT), 2018, pp. 1-5, doi: 10.1109/EAIT.2018.8470438.

[14] P. Y. Simard, D. Steinkraus and J. C. Platt, "Best practices for convolutional neural networks applied to visual document analysis," Seventh International Conference on Document Analysis and Recognition, 2003. Proceedings., 2003, pp. 958-963, doi: 10.1109/ICDAR.2003.1227801.

[15] S. Tayeb et al., "Toward data quality analytics in signature verification using a convolutional neural network," 2017 IEEE International Conference on Big Data (Big Data), 2017, pp. 2644-2651, doi: 10.1109/BigData.2017.8258225.

[16] K. Amrutha and P. Prabu, "ML Based Sign Language Recognition System," 2021 International Conference on Innovative Trends in Information Technology (ICITIIT), 2021, pp. 1-6, doi: 10.1109/ICITIIT51526.2021.9399594.

[17] A. Er-Rady, R. Faizi, R. O. H. Thami and H. Housni, "Automatic sign language recognition: A survey," 2017 International Conference on Advanced Technologies for Signal and Image Processing (ATSIP), 2017, pp. 1-7, doi: 10.1109/ATSIP.2017.8075561.

[18] A. Kumar, K. Thankachan and M. M. Dominic, "Sign language recognition," 2016 3rd International Conference on Recent Advances in Information Technology (RAIT), 2016, pp. 422-428, doi: 10.1109/RAIT.2016.7507939.

[19] Y. Zhao, W. Wang and Y. Wang, "A real-time hand gesture recognition method," 2011 International Conference on Electronics, Communications and Control (ICECC), 2011, pp. 2475-2478, doi: 10.1109/ICECC.2011.6066597.

[20] H. -S. Park and K. -H. Jo, "Real-time hand gesture recognition for augmented screen using average background and camshift," The 19th Korea-Japan Joint Workshop on Frontiers of Computer Vision, 2013, pp. 18-21, doi: 10.1109/FCV.2013.6485452.

[21] S. Albawi, T. A. Mohammed and S. Al-Zawi, "Understanding of a convolutional neural network," 2017 International Conference on Engineering and Technology (ICET), 2017, pp. 1-6, doi: 10.1109/ICEngTechnol.2017.8308186.

[22] R. Chauhan, K. K. Ghanshala and R. C. Joshi, "Convolutional Neural Network (CNN) for Image Detection and Recognition," 2018 First International Conference on Secure Cyber Computing and Communication (ICSCCC), 2018, pp. 278-282, doi: 10.1109/ICSCCC.2018.8703316.

[23] J. Yang and J. Li, "Application of deep convolution neural network," 2017 14th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), 2017, pp. 229-232, doi: 10.1109/ICCWAMTIP.2017.8301485.

[24] Y. -S. Ting, Y. -F. Teng and T. -D. Chiueh, "Batch Normalization Processor Design for Convolution Neural Network Training and Inference," 2021 IEEE International Symposium on Circuits and Systems (ISCAS), 2021, pp. 1-4, doi: 10.1109/ISCAS51556.2021.9401434.

[25] T. Sledevic, "Adaptation of Convolution and Batch Normalization Layer for CNN Implementation on FPGA," 2019 Open Conference of Electrical, Electronic and Information Sciences (eStream), 2019, pp. 1-4, doi: 10.1109/eStream.2019.8732160.

[26] X. Wang, Z. Zhou, Z. Yang, Y. Liu and C. Peng, "Spatio-temporal analysis and prediction of cellular traffic in metropolis," 2017 IEEE 25th International Conference on Network Protocols (ICNP), 2017, pp. 1-10, doi: 10.1109/ICNP.2017.8117559.

[27] S. Umakanthan, S. Denman, S. Sridharan, C. Fookes and T. Wark, "Spatio Temporal Feature Evaluation for Action Recognition," 2012 International Conference on Digital Image Computing Techniques and Applications (DICTA), 2012, pp. 1-8, doi: 10.1109/DICTA.2012.6411720.

[28] Y. Srivastava, V. Murali and S. R. Dubey, "PSNet: Parametric Sigmoid Norm Based CNN for Face Recognition," 2019 IEEE Conference on Information and Communication Technology, 2019, pp. 1-4, doi: 10.1109/CICT48419.2019.9066169.

[29] H. Kazama and S. Tsukiyama, "Gaussian Mixture Reduction Methods Using Support Vector Machine," TENCON 2019 - 2019 IEEE Region 10 Conference (TENCON), 2019, pp. 411-416, doi: 10.1109/TENCON.2019.8929299.

[30] F. Zhan, "Hand Gesture Recognition with Convolution Neural Networks," 2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI), 2019, pp. 295-298, doi: 10.1109/IRI.2019.00054.