Hand Gesture Recognition

Advanced Machine Learning Final Paper Presentation

- Sandhya Cheepurupalli

Abstract:

Hand gesture recognition (HGR) is one of the most actively researched areas in human-computer interaction. Although several hand gesture recognition modalities have been investigated over the past three decades, recent years have seen a resurgence in the field thanks to hardware advancements and deep learning algorithms. This research paper proposes a compact convolutional neural network (CNN) model that, compared to current CNN designs, delivers improved classification accuracy with fewer parameters for recognizing gesture-based sign language. When performance is examined, the proposed model outperforms the VGG-11 and VGG-16 architectures and other cutting-edge methods. Additionally, it is discovered that the suggested model is invariant to scaling and rotation. Overall, the suggested model is a promising method for understanding sign language since it has the potential to identify the most motions with the least number of mistakes.

Keywords: hand gesture recognition, sign language recognition, deep learning, feature extraction, classification

Introduction

Human existence and the advancement of civilization have depended heavily on communication. It has shaped our social characteristics, from our earliest gestures to the emergence of formal languages. To reflect the variety of ways we communicate today, there are thousands of spoken languages and hundreds of sign languages. More than 6,700 spoken languages and several hundred sign languages highlight the world's extensive linguistic diversity [1]. Human relationships depend on communication, and those who struggle to express themselves may feel alone and misunderstood. People who have speech problems, hearing problems, or are deaf deal with these difficulties daily. They have relied heavily on hand gestures as a mode of communication, although it might be difficult for others who need help understanding hand gestures to comprehend them. As a result, the translation of hand gestures into text has been suggested as a possible fix.

Research on contactless hand gesture detection systems has recently been conducted employing cutting-edge modalities like radio frequency (RF), ultrasound, and computer vision (CV). Google's Project Soli, which uses a millimeter wave radar on-chip operating at 60 GHz, is an amazing illustration of this development. The ability of this ground-breaking technology to recognize minute finger and hand movements opens fascinating prospects for improved human-machine interaction and clear communication interfaces [2].

While RF-based and sound-based modalities have improved performance and reliability for specific applications, they still face challenges in regular use cases with varying environmental parameters. As a result, there has been a significant focus on computer vision-based gesture detection systems, thanks to advancements in Artificial Intelligence (AI) and camera technology. In recent years, computer vision has become the most extensively explored field of research for hand tracking and gesture recognition. Various techniques, including skin color detection, appearance detection, motion-based detection, skeleton-based detection, and depth detection, have been employed to achieve accurate and robust hand gesture recognition using computer vision. These advancements promise to create more dependable and versatile gesture recognition systems in diverse real-world scenarios.

Static and dynamic hand movements are two different categories in sign language. Continual hand movements characterize dynamic gestures instead of static gestures, which comprise fixed hand and finger locations over time. Vision and contact-based methods can be used to recognize these hand gestures for sign language translation.

The signer must wear electronic circuitry, such as data gloves, bracelets, or accelerometers, to use the contact-based method. These tools track variations in movement and send the data to a computer for further analysis. Although this method has produced encouraging results, it can be costly and inconvenient for frequent human-computer contact.

On the other hand, the vision-based approach is more user-friendly because it collects the signer's data using a camera. Using image processing algorithms in this method lessens the reliance on sensory apparatuses for data analysis. To translate sign language, this research suggests a vision-based technique for recognizing static hand motions. The suggested method uses a cutting-edge, reliable convolutional neural network (CNN) methodology. The VGG-11 and VGG-16 architectures are also modified and implemented to recognize gestures.

The proposed model in this research demonstrates several advantages over existing state-of-the-art approaches, including high accuracy and reduced training time. The major contributions and novelties of this paper are as follows:

- Introduction of a novel and robust model called G-CNN (Gesture Convolutional Neural Network) for hand gesture recognition.
- Collection of a comprehensive dataset consisting of 43 classes of hand gestures.
- Empirical fine-tuning of hyper-parameters such as kernel width, epochs, batch size, and learning rate to ensure efficient model training.
- Thorough experimental analysis conducted using various evaluation metrics, including accuracy, loss, the recognition accuracy of each class, and training time. The results are compared with VGG-11 and VGG-16 models using the same dataset.
- Demonstration of the generalization ability of the proposed model through the evaluation of an augmented dataset. The model shows competent results and exhibits robustness to rotation and scaling transformations.
- Evaluation of the model's performance using 10-fold cross-validation to assess its consistency and reliability.
- Remarkable results were achieved across all evaluation metrics, indicating the model's capability to handle complexity and hand occlusion.

Related Work

Researchers have devoted much time to studying gesture recognition for sign language, leading to the development of contact-based methods. Using LSTM for classification, [3] used a wearable device with six inertial measuring units (IMUs) to recognize 28 words in American Sign Language (ASL). To analyze skeleton sequences for bidirectional communication, [4] suggested a gesture recognition method for Chinese sign language utilizing recurrent neural networks (RNN), extracting hand position and finger movements, and using LSTM for classification.

To translate sign language, researchers have proposed various vision-based methods for hand motion detection. In their method, [5]used edge detection and grey-scale transformation to extract features, then gesture recognition and text display using template matching. Hand gesture recognition using YCbCr skin segmentation, Zernike moments for feature extraction, and SVM for classification were proposed by [6] These vision-based methods illustrate various methods for understanding sign language hand gestures, providing prospective solutions for efficient communication.

Gesture Recognition Method

Figure 1 illustrates the framework of the model used in this paper.

1. Data Collection:

In this paper, we used a pre-available dataset from Kaggle, which is readily processed. Next, images have been fed into the proposed model of CNN for feature learning, and the classification has been performed using the Softmax classifier.

2. Data Pre-processing:

Data pre-processing is done to improve the quality of the dataset. As each of the collected sign gestures in the dataset was of different sizes and very high resolution, only the hand gestures have been cropped. Then to make the dataset used for the machine learning model, each image is spatially down-sampled to the size of 256 × 256. This reduced image size and resolution minimized the computation complexity and helped in the faster convergence of the classifier.

3. Data Labelling:

For supervised learning algorithms, data labeling is essential for preparing datasets. It entails giving relevant tags or labels to data samples to lay the groundwork for classifier learning. For this article, sign images were gathered and divided into 43 classes, each with a designated folder with photos exclusive to that class. As a result, the data were labeled according to their respective class names, which allowed the classification model to associate with and learn from the labeled data efficiently.

4. Proposed Model and Architecture [7]:

Figure 2 shows the architecture of the G-CNN model used in the paper. A CNN-based model created exclusively for gesture-based sign language recognition is used in this investigation. Gesture CNN (G-CNN) is the suggested model's name, which has 12 levels: 4 convolutional layers, three pooling layers, two dropout layers, two fully connected layers, and one softmax layer. Instead of the other CNN architecture based on huge filter sizes, the weighted layer uses small filters of 3, 2, and 1.

The input gesture images—256x256 pixels—undergo several processing phases in the proposed gesture identification system. First, a convolutional layer is used, where filters are applied to the input image to extract features. These filters' weights are automatically learned and updated during the feature extraction. 32 3x3x32 dimensional convolutional filters are used, which leads to the extraction of 32 high-level features represented by a dimension of 256x256x32. After the convolutional layer, a hyperbolic tangent activation function adds nonlinearity and teaches nonlinear decision limits. The suggested convolutional neural network (CNN) architecture is not very deep, but the system performance is not adversely affected by the usage of the tanh activation function. As shown in the results section, it helps to speed up model training. As a result, it is thought that using the tanh function in this task is helpful. Algorithm 1 further explains the significance of the activation function, and Equation (1) provides a mathematical expression for the tanh function.

$$f(x) = \frac{1-\exp\ ^{\wedge}-2x}{1+\exp\ ^{\wedge}-2x} \quad (1)$$

Multiple convolutional and max-pooling layers are stacked to create the spatiotemporal representation of gestures. Using the max-pooling process, the resulting feature maps are downscaled by a factor of two. The model comprises a hyperbolic tangent (tanh) activation function and four convolutional layers with a stride of 1. Each convolutional layer's associated depth is 32, 64, 64, and 128, and its kernel sizes are 3, 3, 1, and 3. Due to the small kernel size, the model can capture even the tiniest texture details in the motions. The feature maps are compressed using max-pooling with a stride of 2 and a filter size of 2.

After the convolutional and max-pooling layers, a group of fully connected layers is used to connect the extracted features for classification. The hidden units in the completely connected layers have sizes of 512 and 84. Two dropout layers are included with probabilities of 0.3 and 0.2 to prevent overfitting during training. These layers randomly remove inactive neurons from the

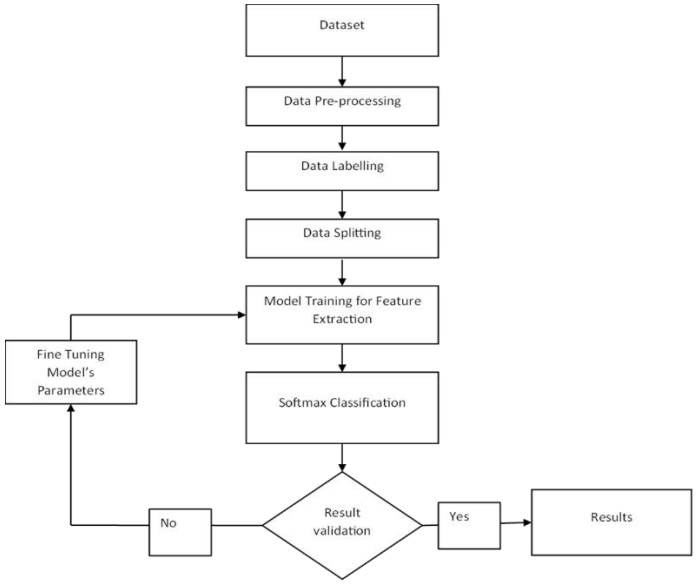

Figure 1: Framework of the Hand Gesture Recognition Model

network. The output of the final fully connected layer is then transmitted through a softmax layer, which uses Equation (2) to compute the probability distribution of the classes and predict the gesture classes.

$$P(y = i|x) = \frac{e^{x^T w_i}}{\sum_{k=1}^{K} e^{x^T w_k}} \quad (2)$$

The model configuration is presented in Table 1, where x^T represents the T-th element of the array and K represents the total count of elements in the array x.

Table 1

Configuration of the proposed G-CNN

Layer Type	No. of Filter	Feature Map Size	Kernel Size	Stride Used
Input Image Layer	-	256 × 256	-	
Convolution 1	32	256 × 256 × 32	3 × 3	1 × 1
Max-pooling 1	1	128 × 128 × 32	2 × 2	2 × 2
Convolution 2	64	128 × 128 × 64	3 × 3	1 × 1
Convolution 3	64	128 × 128 × 64	1 × 1	1 × 1
Max-pooling 2	1	64 × 64 × 64	2 × 2	2 × 2
Convolution 4	128	64 × 64 × 128	3 × 3	1 × 1
Max-pooling 3	1	32 × 32 × 128	2 × 2	2 × 2
Dropout1				
Fully connected 1		512 × 1		
Fully connected 2		84 × 1		
Dropout2				
Output Layer		43 × 1		

The advantages of the G-CNN proposed model are as follows:

- Automatic Feature Extraction: The G-CNN model can automatically extract key features from input frames. This eliminates the need for manual feature extraction, which is required in feature-based recognition systems. The G-CNN model performs better by automatically learning relevant features e than feature-based approaches.
- Compact Architecture: The G-CNN model has a compact architecture consisting of 4 convolutional layers, 3 pooling layers, 2 dropout layers, 2 fully connected layers, and 1 softmax layer. This compact representation results in fewer trainable parameters and a lower computational load. This is particularly advantageous for real-time applications where computational efficiency is crucial.

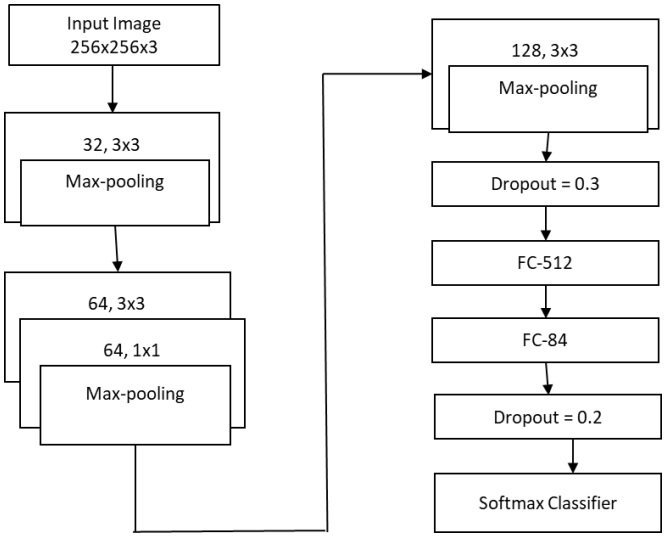

Figure 2: Architecture of G-CNN Model

- Efficient Training: Despite its less deep architecture, the G-CNN model achieves good recognition results while consuming less training time than state-of-the-art deep learning models. This indicates that the G-CNN model balances model complexity and training efficiency, making it a practical choice for gesture recognition tasks.

5. VGG-11 and VGG-16:

Convolutional neural networks (CNNs) have come a long way since [8] first suggested the VGG network architecture in 2014. Its main goal was to investigate the impact of greater network depth on classification accuracy. Since its debut, VGG models have performed astoundingly in various computer vision tasks, including texture identification, object detection, and picture captioning [9]; [10].

Nevertheless, despite their broad popularity, more is needed about how VGG models might be used to recognize hand gestures, particularly in sign language. The effectiveness of VGG models for identifying fingerspelled and isolated words in sign language translation systems has yet to be adequately explored in prior research. Therefore, it is still being determined if VGG models can attain high classification accuracy in such systems.

In the context of sign language, this research seeks to assess the performance of the VGG-11 and VGG-16 models in hand gesture recognition. The obtained dataset for this investigation has been scaled down to [256 256], whereas the original VGG models were created and tested with input photos of size [224 224 3]. As a result, the VGG-11 and VGG-16 models are adjusted appropriately to handle this prepared dataset correctly.

There are 11 weighted layers altogether in the VGG-11 model, comprising 8 convolutional layers, 3 fully connected layers, and one Softmax layer. The pooling layers use a filter size of [2 2] with a stride of 2, while each convolutional layer in VGG-11 has a filter size of [3 3] with a stride of 1. Similarly, the VGG-16 model has 16 weighted layers, including 13 convolutional layers, three fully connected layers, and a Softmax layer at the bottom.

The suggested model and the VGG-11 and VGG-16 architectures are trained and tested on various sign language datasets in this work, along with other CNN

models. Each class's images are randomly selected and split into three sets: 70% are used for training, 10% for validation, and 20% are used for testing. 60 training iterations are performed on the network, with a batch size of 32 samples for each training step. The Adadelta optimizer, which can adjust learning rates based on gradient updates, is used. The Adadelta optimizer's starting learning rate and decay factor are 1 and 0.95, respectively.

Table 2

Classification results

MODEL	G-CNN	VGG-11	VGG-16
ACCURACY (%)	94.83	93.6	93
LOSS	0.351	0.463	0.491

Experimental Analysis

The performance of the suggested CNN model (G-CNN), two CNN architectures (VGG-11 and VGG-16), and sign language recognition are examined in this article. This section explains the experimental findings for these models, and the table below offers a comparison with other cutting-edge methods. Metrics like accuracy, loss, processing time, and classification prediction outcomes are used to assess their performance.

1. Training Time

Table 3

Training time and total trainable time

MODEL	TIME UTILIZED (IN MINS)	PARAMETERS USED
G-CNN	9:21	67, 250, 830
VGG-11	40:15	160, 297, 362
VGG-16	44:33	165, 791, 570

2. Classification Prediction Result

The research also calculates a performance metric called the confusion matrix to improve the evaluation of the suggested technique. This matrix gives an overview of samples that were successfully and wrongly predicted for each class, making it possible to determine the recognition accuracy for each class separately. The chart shows the three CNN models' detailed recognition accuracy for each class in Dataset. The findings show that the suggested G-CNN performs admirably in most courses.

3. Other Parameters

In real-time applications, computational time is crucial in hand gesture recognition. Table 6 presents the training time required by three different CNN architectures. Additionally, the table provides details about the parameters used by these models, enabling the assessment of their complexity. The total number of trainable parameters can be calculated using the provided expressions.

The total parameters P_{conv} for each convolutional layer can be calculated with Eq. (5).

$$P_{conv} = (width_{filter} * height_{filter} * no\ of\ filters\ of\ previous\ layer + 1) * no\ of\ filters \quad (5)$$

The total parameters for each Fully connected layer (Pfc) can be calculated with Eq. (6).

10-fold cross-validation results of proposed work (G-CNN) for the dataset

$$Pfc = ((currentlayerc * previouslayerp) + 1 * c) \quad (6)$$

The results unequivocally show that the proposed GCNN architecture uses fewer parameters and takes less time to compute than the sophisticated CNN models. This exhibits the suggested architecture's efficiency and efficacy regarding computational resources.

4. 10-fold Cross Validation

The performance of the G-NN model throughout the full dataset has been evaluated and stabilized using the 10-fold cross-validation technique. Table 7 thoroughly analyzes the model's performance across various subsets of data by presenting the assessment results for each tenfold.

5. Comparison of the classification result

In this section, a comprehensive comparison between the proposed G-CNN method and existing methods in the field of hand gesture recognition is presented.

The G-CNN model surpasses all these methods by achieving the highest accuracy of 94.83%.

K-FOLD	K=1	K=2	K=3	K=4	K=5	K=6	K=7	K=8	K=9	K=20
ACCURACY	94.83	94.25	94.68	93.18	94.35	95.26	92.78	93.68	94.25	94.87

Conclusion

In this paper, a deep learning-based method for hand gesture identification is presented. The suggested method for categorizing hand gestures combines a small, effective G-CNN model with modified VGG11 and VGG-16 architectures. This vision-based model's capacity to do away with user dependence and the requirement for external gear makes it useful for real-world applications.

The research contributes substantially by outperforming current state-of-the-art methods and attaining good recognition results. The proposed G-CNN model has the highest classification accuracy, with 94.83%. The model's resilience to rotation and scaling transformations is further demonstrated by testing it with enriched data and analyzing several efficiency indicators.

The comparative analysis demonstrates the suggested model's higher performance in classifying 43 different gestures with a low error rate. The deep learning architectures can be further tuned in future work to improve hand gesture identification and perform more thorough comparisons. By investigating and improving these structures, it may be possible to reduce mistake rates in real-time sign language recognition.

References

[1] U. &. P. N. Zeshan, "Typology of sign languages. Camb. Handb. Linguist. Typology 1–33 (2017)".

[2] J. e. a. S. Lien, " Ubiquitous gesture sensing with millimeter wave radar. ACM Trans. Graph. TOG 35, 1–19 (2016)".

[3] T. W. &. K. B. J. (. Chong, "American sign language recognition system using wearable sensors with the deep learning approach. The Journal of the Korea Institute of electronic communication sciences, 15(2), 291–298.".

[4] Q. Q. M. &. Y. Y. (. Xiao, "Skeleton-based Chinese sign language recognition and generation for bidirectional communication between deaf and hearing people. Neural Networks, 125, 41–55.".

[5] S. &. B. M. M. (. Shrenika, " In Sign Language Recognition Using Template Matching Technique (pp. 1–5). IEEE.".

[6] P. K. S. C. J. &. L. A. (. Athira, "A signer independent sign language recognition with co-articulation elimination from live videos: an indian scenario. Journal of King Saud University-Computer and Information Sciences.".

[7] S. S. Sakshi Sharma, "Vision-based hand gesture recognition using deep learning for the interpretation of sign language," *Expert Systems with Applications.*

[8] K. &. Z. A. (. Simonyan, "Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556".

[9] Y. Z. C. W. J. S. M. &. Z. X. (. He, "Bounding boX regression with uncertainty for accurate object detection. In In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (pp. 2888–2897).".

[10] Y. S. P. W. N. &. S. Y. (. Liu, "A survey and performance evaluation of deep learning methods for small object Detection. Expert Systems with Applications, 114602".

[11] U. &. P. N. T. o. s. l. C. H. L. Zeshan, pp. 1-33, 2017.