A Method of Hand Gestures Recognition using Convolutional Neural Network

Ragapriya Saravanan, Sindhu Retnaswamy, and Shirley Selvan

Abstract The progress in the realm of machine learning has caused the field of image processing to undergo a significant advancement. Image processing is a vast subject due to the diverse algorithms and techniques that can be implemented. Sign Language is the hearing disabled community's way of conveying information and this dissertation was drafted to understand and provide them with what little assistance we can. Visual cues and signs are used to convey messages and intentions of the speaker. It is a well-developed language that has its own vocabulary and grammar. In this paper, we analyze the various methods used in converting Sign Language into text which can be read or to audio that can be heard. This treatise also includes our very own methodology, which makes use of a CNN architecture called AlexNet and discusses the results of the same.

Keywords Sign Language · ASL · Fingerspelling · Pre-processing · Feature extraction · Recognition · Methodology · Algorithm · Dataset · Neural network

1 Introduction

Sign Language is a way of communication, rather a way of life practiced by the hearing impaired. It has existed among the hearing disabled community since ancient times. The spoken language that uses the reverberations from the mouth and is comprehended by the ear is difficult for them to understand. Sign Languages were established to exploit the unique features of visual modality (sight), along with tactile features (touch). Spoken Language is linear in nature as in, only one sound can be made or received at a time, whereas Sign Language, on other the hand, is visual and can therefore use a simultaneous expression through visual perception. Sign Language varies from each hearing disabled community and the place they live

R. Saravanan (✉) · S. Retnaswamy · S. Selvan
Department of Electronics and Communication Engineering, St. Joseph's College of Engineering, Chennai, India

A. Sivasubramanian et al. (eds.), *Futuristic Communication and Network Technologies*, Lecture Notes in Electrical Engineering 792,
https://doi.org/10.1007/978-981-16-4625-6_34

in; hence, Sign Languages are not universal. Linguists have perceived and distinguished 137 Sign Languages to date including American Sign Language, British Sign Language, Indo—Pakistani Sign Language, etc.

American Sign Language (ASL) is a language that is completely different and separate from the English language (see Fig. 1). ASL contains all the fundamental features of the language, with its own rules for pronunciation, word formation, and order. Parents are the main source of a child's early acquisition of language, but for children who are hearing impaired and dumb, additional people are necessary as models for language acquisition. A child who has a hearing impairment, born to parents with a similar impairment, naturally picks up Sign Language as a hearing abled child acquires spoken language from hearing abled parents.

A few methods of hand sign recognition from different authors over the years have been elucidated in the literature. A simple rule classifier to predict gestures is proposed [1]. Background subtraction is applied to the captured images to separate the hand region, which is the region of interest in this case from the background. After the hand is detected, the fingers and the palm region are segmented separately with the help of 3 parameters, which include the palm point [2], inner circle of

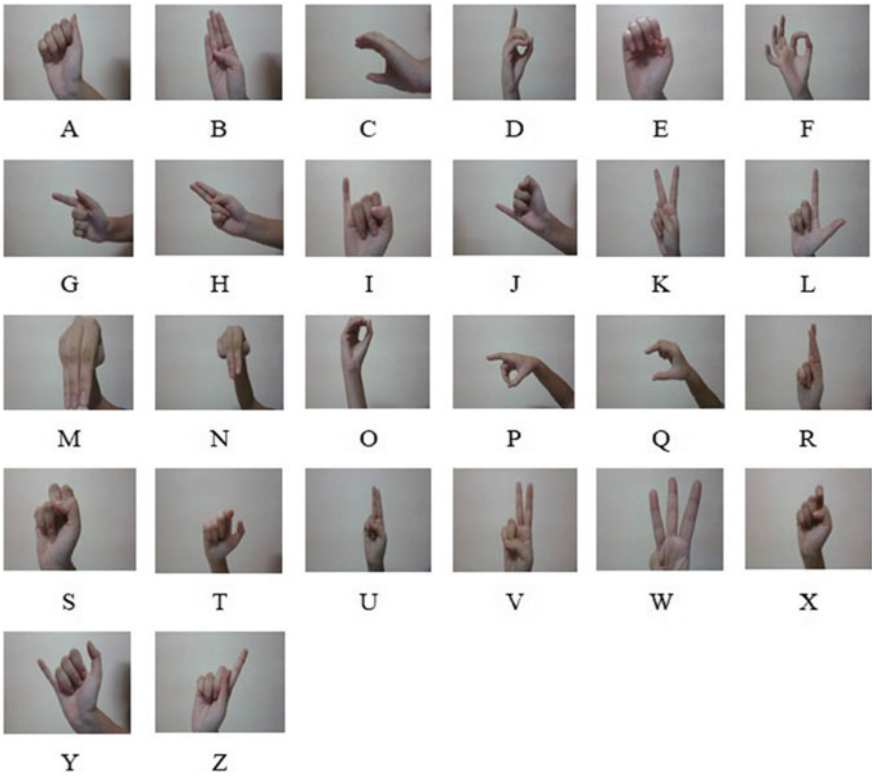


Fig. 1 Alphabets in ASL

maximal radius, and the wrist points and palm mask. When the fingers are detected and recognized, a simple rule classifier can be used to classify the hand gesture. In the rule classifier, the hand gesture is predicted according to the number and type of fingers detected. This classifier has high accuracy and is quick in detecting the distance transform to identify the sections of the hand such as the palm and the fingers. Despite having high accuracy, this method does not predict the gesture as well as other methods [3].

The speed up robust features (SURF) algorithm operates on still images of hand gestures shown in front of the webcam created with the help of frame capture function from real-time video capture. The points of interest of the grayscale images are detected using the SURF algorithm [4]. These points of interest from the captured images as well as the reference images are stored in the database. Following this, features are extracted and selected by SURF. The final step would be the recognition of the hand gestures by matching the reference features and reference points of the reference image with that of the input image. The use of box filters in the SURF algorithm provides the benefit of real-time applications such as tracking and object recognition. It provides robustness to rotation and also performs well with scale illumination changes. But the problems associated with SURF include its difficulty to track edges, dependency on view, and fragility [5].

An approach that uses the scale invariant feature transform (SIFT) algorithm was put forward in which a video of the hand is captured through a live camera using OpenCV. The SIFT algorithm detects key points or interest points of an object to provide a unique feature descriptor for the object, which are grouped to form numerous feature vectors for each image. The successive step would be orientation detection which will consider the movement of the hand. When the entire algorithm has completed compiling, the set of images fed as input are then translated into a one-dimensional array containing the characters corresponding to the alphabet shown in the image. When the alphabet is recognized, using speech conversion, the recognized text is converted to speech, and audio output is executed. The major interest in using the SIFT algorithm is that the feature vectors are not affected by object scaling or rotation. Also, the usage of multiple feature vectors would provide reasonable accuracy. The demerits of SIFT are that it is slow, mathematically complicated, and not good at illumination changes [11].

Another method of gesture recognition using a few segmentation techniques and convolutional neural network (CNN) is proposed. Once the hand is segmented, a morphological operation [12] is applied to the images to remove holes and noise present in both the object and the background. An arbitrary color model is then provided as input for the CNN architecture. Finally, based on the feature maps computed by the previous layers, the multilayer perceptron neural network performs classification. The architecture of the neural network and the provided training set influence the performance of the model greatly. There may be a wrong assignment of labels due to overfitting which is a major hurdle [13].

The paper is framed in the following way: Sect. 2 highlights the course of action which was developed for training a model using AlexNet [14] to recognize the signs of the ASL alphabets in a detailed manner. The trained model is tested with live

frames captured through the webcam of the laptop and the results obtained for the same are discussed in Sect. 3. Finally, the conclusion is drafted and the means to improve the outcome along with the future scope of the project are enunciated in Sect. 4.

2 Methodology

We propose a method that uses AlexNet to recognize hand gestures. AlexNet is one of the architectures of CNN, designed by Alex Krizhevsky, published along with Ilya Sutskever, and Geoffrey Hinton. It is an incredibly powerful architecture that is adept at achieving high accuracies on very difficult datasets.

AlexNet has around 60 million parameters and 650,000 neurons making the network more accurate and intricate. AlexNet is suitable for object detection and many other applications in computer vision and image processing [15]. AlexNet architecture is 8 layers deep, comprising five convolutional layers and three fully connected layers.

2.1 Support Package

The pre-trained version of the network can be downloaded in the Add-On Explorer of the MATLAB software under the name "Deep Learning Toolbox Model for AlexNet Network". The pre-trained network is capable of classifying images into 1000 object categories, such as the keyboard, mouse, pencil, etc., as it was trained for 6 days using 2 powerful GPUs (see Fig. 2).

(a). Switch **(b).**Window Shade **(c).** Keyboard **(d).** Mouse

Fig. 2 Examples of classification by pre-trained network

2.2 Dataset

The dataset consists of a directory of 26 folders, each folder representing an alphabet of ASL, taken from Kaggle. Every folder contains a collection of 110 RGB images of dimensions 200 × 200 pixels, captured with a white background.

2.3 Steps for Recognizing Hand Gestures Using AlexNet

The steps for identifying hand gestures using images and videos with regard to AlexNet are presented below (see Fig. 3.).

Modify dataset. The input images fed to the AlexNet network should be RGB images of size 227 × 227 pixels. The dataset is modified and resized to fit this criterion.

Load dataset. The path to the dataset is fixed and the resized images are loaded onto the MATLAB workspace. The dataset is then split in such a way that 80% of the

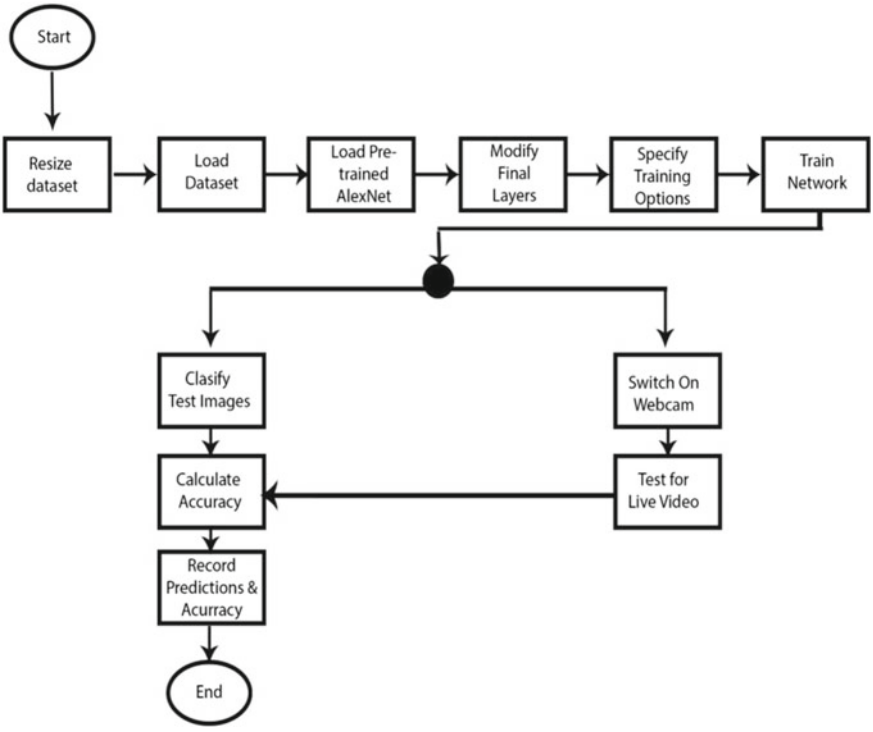


Fig. 3 Process flow diagram

Table 1 Dataset division

Dataset (for each alphabet)	Training images (for each alphabet)	Testing images (for each alphabet)
110	88	22

dataset is grouped into training images and the remaining 20% into testing images as indicated in Table 1.

Load AlexNet. The pre-trained AlexNet neural network model from the support package is then loaded onto the file.

Modify Final Layers. In order to fit the AlexNet to our custom specifications, the final layers are to be modified accordingly, which includes the classification layer and softmax. The basic architecture of AlexNet with 25 layers is used (see Fig. 4.).

Train Network. The network model is now trained with the input images from the dataset, the network architecture of AlexNet, and the training options. The training options specified include the optimizer: stochastic gradient descent with momentum (*sgdm*), initial learning rate: 0.001, maximum number of epochs: 10, and the mini batch size: 64.

Classify Validation Images. The trained model is then allowed to classify the test images. It assigns each image, a label representing the predicted output.

Calculate accuracy. Using the predicted output from the test images, the accuracy is calculated.

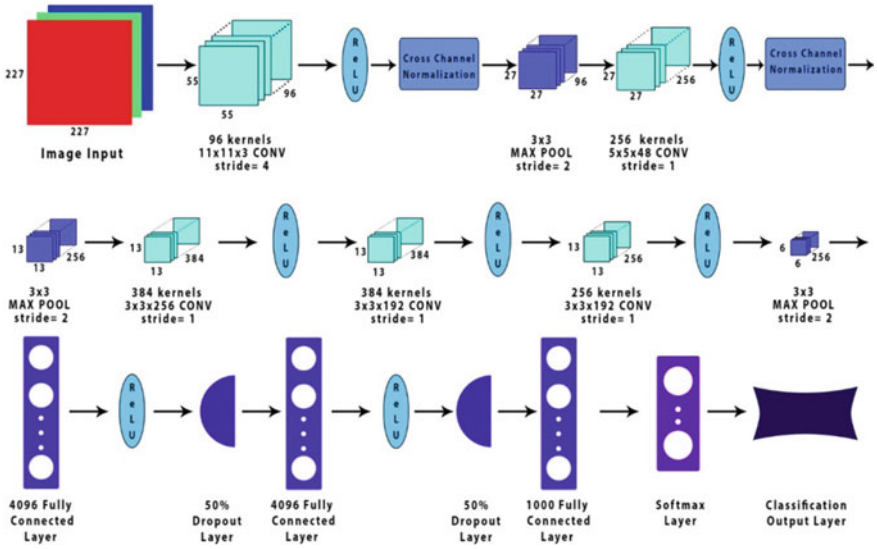
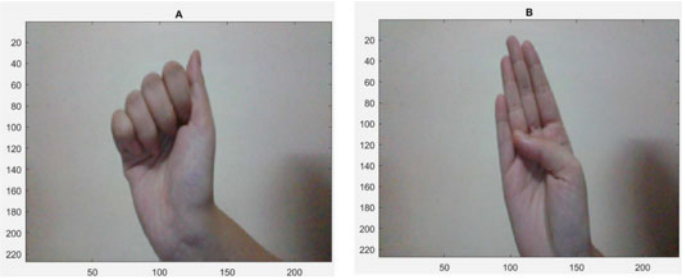
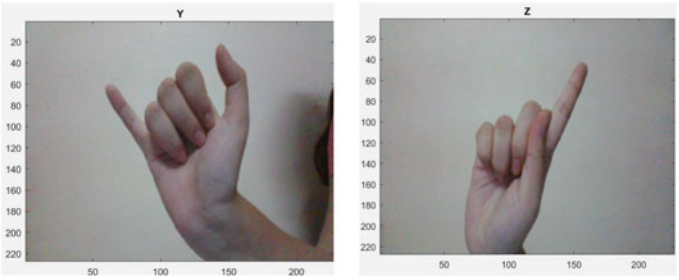


Fig. 4 Architecture of AlexNet to be modified for our methodology



(a). Output for signs A and B



(b). Output for signs Y and Z

Fig. 5 Examples of correct prediction from live video

Live Testing. The webcam of the laptop is used to capture RGB images of the hand signs by taking individual frames from the video at particular intervals. The captured images are resized to 227 × 227 pixels, which is then classified using the trained network. The predicted output is displayed as the title of the image (see Fig. 5).

Problems Encountered During Training. This approach was executed on a machine that accommodated a GPU that proved to be incapable of training the AlexNet model for more than 10 epochs. In the case of using a sizeable dataset or increasing the number of iterations, the machine was unable to cope and halted before the completion of the training. Hence, it is recommended to use a fairly powerful GPU while training this model to get better results and accuracy.

3 Results and Discussion

The particulars of the results acquired are specified in Table 2. The accuracy obtained during the training of the AlexNet model using an *sgdm* optimizer for 10 epochs was 100% while testing using the test images. The same was tested and an accuracy of 76.92% was obtained for live images of hand signs captured by the laptop's webcam.

Table 2 Performance analysis of proposed method

Architecture used	Training time (min)	Test images		Live video	
		Correct prediction (alphabets)	Accuracy	Correct prediction (alphabets)	Accuracy
AlexNet	15	26/26	100%	20/26	76.92%

This model was unable to recognize alphabets R, S, T, U, V, and X for live images from the video. This variation in accuracy for test images and live video can be attested to any of the following reasons below:

- The similarity of the descriptor vectors of the signs.
- Inadequate time given for the model to train due to machine restrictions.

Limitations of Proposed Method. The high accuracy of any system that uses AlexNet can be attributed to the multiple layers present in its architecture but removing even a single layer leads to a 2% loss. Thus, the depth is important for the accuracy of AlexNet. It is also computationally expensive since it contains 3 fully connected layers.

4 Conclusion

This work explains how the ASL alphabet hand signs are captured using a webcam and converted into text. For the conversion of gesture to text, a CNN architecture called AlexNet is used. AlexNet is an architecture of CNN that has good accuracy and training speed. It avoids issues such as overfitting, making the network robust to errors and misrecognition. The proposed architecture can be used to determine ASL alphabet hand signs easily by using the trained model in a classification program.

Further work can be done on this approach by increasing the total dataset, as the model was trained only on 110 images for each alphabet for 10 epochs. It can also be developed to recognize words of ASL and non-stationary signs.

References

1. Patel R, Dhakad J, Desai K, Gupta T, Correia S (2018) Hand gesture recognition system using convolutional neural networks. In: 2018 4th international conference on computing communication and automation (ICCCA), pp 1–6
2. Gautam AK, Department of Electronics and Communication Engineering, Delhi Technological University Delhi, India, Kaushik A, Department of Computer Science and Engineering, Kurukshetra University, Haryana, India (2017) American Sign language recognition system using image processing method. Int J Comput Sci Eng (IJCSE)

3. Zhao Shan Chen Z-H, Kim J-T, Liang J, Zhang J, Yuan Y-B (2014) Real-time hand gesture 168 recognition using finger segmentation. The Scientific World Journal, Hindawi Publishing 169 Corporation. https://doi.org/10.1155/2014/267872
4. Soniya M, Sarah Suhasini P (2019) Integrated SURF and spatial augmented color feature based Bovw model with Svm for image classification. Int J Eng Adv Technol (IJEAT) 8(6). ISSN: 2249–8958
5. Kour K, Mathew L (2017) Sign language recognition using image processing. Int J Adv Res Comput Sci Softw Eng 7(142). https://doi.org/10.23956/ijarcsse.v7i8.41
6. Gaikwad S, Shetty A, Satam A, Rathod M, Shah P, Department of computer engineering, MCT ”s rajiv gandhi institute of technology, Mumbai, India (2019) Recognition of American Sign language using image processing and machine learning. Int J Comput Sci Mobile Comput (IJCSMC)
7. Kumar RS, Srivastava M, Computer science and engineering department Jaypee University Anoopshahr, Patna, India (2019) Hand Gesture recognition using image analysis and neural network. Int J Res Advent Technol Special Issue
8. Pinto RF, Borges CDB, Almeida AMA, Paula IC, Universidade Federal do Ceará, Sobral, Ceará 62010–560, Brazil (2019) Static hand gesture recognition based on convolutional neural networks. J Electr Comput Eng
9. Krizhevsky A, Sutskever I, Hinton GE (2012) ImageNet classification with deep convolutional neural networks. In: Pereira F, Burges CJC, Bottou L, Weinberger KQ (eds) Advances in neural information processing systems 25. Curran Associates, Inc., pp 1097–1105
10. Sudha KK, Sujatha P (2019) A Qualitative analysis of googlenet and alexnet for fabric defect detection. Int J Recent Technol Eng (IJRTE) ISSN: 2277-3878