Ensemble Image Filtering for Sign Language Classification using Convolutional Neural Networks

Nikunj Lad
College of Engineering
Northeastern University
Boston, USA
lad.n@husky.neu.edu

Johail Sherieff
College of Engineering
Northeastern University
Boston, USA
mehaboobsherieff.j@husky.neu.edu

***Abstract*— This paper presents different image filtering techniques for sign language classification using Convolutional Neural Networks (CNN). The research performed in this paper focuses on trying to analyze the efficiency of CNN having used an ensemble of filters. After initial reshaping and normalizing of images, filtering is performed using image filters like Gabor Filter, Sobel Filter, Watershed algorithm, Adaptive Histogram Equalization and Adaptive Gaussian Thresholding in order to observe their effect on classification accuracy of CNN. The proposed system makes use of 2 sign language image sets – A and B for evaluating classification accuracy. A total of 6000 (3000 for each alphabet) images are used for training the model and 82 (41 for each alphabet) images are used for testing. Independent processing of images using above filters yielded an average accuracy of 80% on test data as compared to 98.78% given by unprocessed CNN's. The application of ensemble of certain filter subsets give an accuracy of 99.20% on the unknown data which is 0.4% improvement on traditional CNN performance.**

Keywords—Gabor Filter, Adaptive Gaussian Thresholding, Adaptive Histogram Equalization, Sobel Filter, Watershed Algorithm, Convolutional Neural Networks

I. Introduction

American Sign Language (ASL) recognition is a heavily yet evolving computer vision problem. Over the past couple of decades, researchers have used different approaches to detect static hand gestures using linear classifiers, neural networks, Artificial Neural Network, Fuzzy Logic, Genetic Algorithm, Hidden Markov Model, Support Vector Machines etc. ASL focuses more on the hand gestures than the facial expressions. The base dataset of ASL includes just the 26 English characters of hand signals. Usually this involves only a single person's hand signs. As a result, developing a robust system can be difficult at considering lack of generalized data. Data consisting of different people of various age groups, belonging to different geographic locations as well as ambient conditions is desirable in order to develop an efficient model having economic value.

Convolutional Neural Networks have been extremely successful in solving the real-world problems of image recognition and classification and were repeatedly implemented to recognize static hand gestures in recent years. In particular, there is work done in the recognition and classification of American Sign Language using deep CNNs, with input-recognition of the hand images to be converted to pixels of the image. The most relevant work to date is L. Pigou et al's application [2] of CNN's to classify 20 Italian Sign Language gestures on a considerably small amount of dataset. Previously, there was only basic camera technology used to generate datasets of images, which excluded depth or contour information while specifically considering pixel values. Attempts at using CNNs to handle the task of classifying images of ASL hand gestures have made some success [3], by using the pre-trained GoogLeNet architecture. While CNN's are highly efficient considering the above cases there are certain places where it fails to perform better. CNN's cannot encode the position and orientation of their object into their predictions. This includes images which has texture information distributed which has certain angular orientations in spatial domain. Another limitation of CNN is it fails to check the relative location of the features compared to each other. The relationship of objects to their surroundings and their relative orientations are something which the CNN's fail to capture from an image.

Filtering is a technique for modifying or enhancing an image by using various preprocessing techniques. Algorithms like Adaptive Histogram Equalization help for feature emphasis, while removal of other features or noises from the data can be performed by Watershed Algorithm. Image processing operations include filtering, smoothing, sharpening, edge enhancement, and detection.

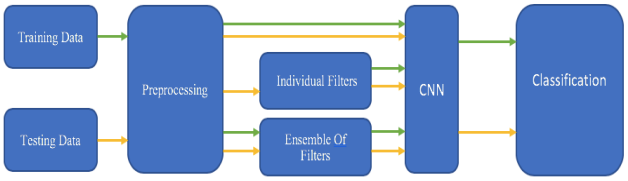
Edge detection is one of the key techniques in image processing for finding the boundaries of objects. It works by detecting discontinuities in brightness between the object pixels and the neighbor pixels. Sobel Filter and Gabor Filters are effective for detecting edges in images along with image textures and their spatial orientations.

This paper focuses on implementing various image filters on images of hand signs in order to observe any increase in classification accuracy of CNN's. Although conventionally, CNN are capable of producing robust results due to their inherent filtering mechanisms, the introduction of an ensemble of filters prior to CNN's architecture for increasing it's classification accuracy is the hypothesis we intend to prove

II. Methodology

A. Block Diagram

Figure 1 represents the pictorial diagram how the project is structured. First step the training and testing data is created. It is then sent to the preprocessing techniques to resize and convert the image to greyscale. Each individual filter are applied to the data and an ensemble of filters are also applied to the data before applying it to the CNN. CNN trains on the training data and then cross validates, where the training data is passed to check how well the model preforms with each individual filter or group of filters.



B. Data Acquisition

In this project, the ASL Alphabet Database is used. This database consists of 26 alphabets where each image is taken under different conditions and orientations by the same user buy proves to be good database. The database consists of 3,000 images of each alphabet that is total of 78,000 images for 26 alphabets with an image size of 200 x 200.

C. Data Processing

Data is resized from higher dimension to a lower dimension to avoid computational cost and increase speed. The images are converted to grayscale values for reducing noise and computational time since a depth of 3 consisting of RGB layers will be more expensive and difficult to model than with images having one layer.

D. Filtering Methods

We have used 5 filtering methods for pre-processing images

1. Gabor Filters

   Gabor filters have been adapted in edge detection and feature extraction. Gabor filters are special classes of the bandpass filter and they allow only a certain band of sinusoidal planes with frequencies and orientation and reject the others which are modulated by a Gaussian envelope. The filter responds well in the edges and texture changes as it distinguishes a particular feature via the filter and the filter has a good distinguishing value at the spatial location of each feature as it can be modeled to find at any angle. If the Gabor filter is oriented at a particular direction and a strong response is given at the locations of the target images that have structures in the same direction of the Gabor filter.

2. Sobel Filters

   The Sobel edge detector is a gradient based method. It works with first order derivatives. It calculates the first derivatives of the image separately for the X and Y axes. The derivatives are only approximations (because the images are not continuous)

   $$G = |G_x| + |G_y|$$

3. Watershed Algorithm

   In the study of image processing, a watershed is a transformation defined on a grayscale image. The name refers metaphorically to a geological watershed, or drainage divide, which separates adjacent drainage basins. The watershed transformation treats the image it operates upon like a topographic map, with the brightness of each point representing its height, and finds the lines that run along the tops of ridges.

   There are different technical definitions of a watershed. In graphs, watershed lines may be defined on the nodes, on the edges, or hybrid lines on both nodes and edges. Watersheds may also be defined in the continuous domain. There are also many different algorithms to compute watersheds. Watershed algorithm is used in image processing primarily for segmentation purposes.

4. Adaptive Histogram Equalization

   Adaptive Histogram Equalization is an image processing technique that handles the contrast adjustment using the image's histogram. It differs from ordinary histogram equalization in respect to the adaptive method computes several histograms for each pixel, it does not increase the contrast for the pixels but uses them to redistribute the lightness values of the image. This method helps in improving the local contrast and therefore enhancing the definitions of edges in the image.

5. Adaptive Gaussian Thresholding

   Adaptive Gaussian Thresholding is an image segmentation technique which works as a non-linear operation that converts a gray-scale image into a binary image as the pixels which are assigned below or above the specified threshold value. If the pixel value is greater than a threshold value, it is assigned to one value (that may be white), else it will be assigned to another value (that may be black). Gaussian Thresholding is an image segmentation technique which works as a non-linear operation that converts a gray-scale image into a binary image as the pixels which are assigned below or above the specified threshold value. If the pixel value is greater than a threshold value, it is assigned to one value (that may be white), else it will be assigned to another value (that may be black).

E. Convolutional Neural Networks

Convolutional Neural Networks (called usually as ConvNets) was initially developed by Yann LeCun back in the 1988 under the name of LeNet. CNN architecture employs a set of layers for feature extraction and for retrieving critical image information from it. It consists of convolutional layers initially where a convolutional matrix traverses across the image using

concept of sliding window. This is done iteratively thereby creating a set of convolutional layers usually in the order of 2. The convolutional matrix is usually odd in size Next, we have the layer having activation function of Rectified Linear Units called generally as ReLu layers. This layer operates using following function:

$$X(z) = \max(0,z)$$

Here the, negative pixel values are rounded up to 0 while the higher values are kept as it is. Further, there is a max pooling layer which helps in reducing the image size by finding the maximum pixel value from a traversing kernel. These layers are repeated in succession in order to create more depth and retrieve more information of image properties. A dropout layer is added in order to avoid overfitting of the data. These layers are then flattened out to be fed to the dense layers. Dense layers are basically multiple hidden layers which are then connected to fully connected layer. The activation function for fully connected layer is different depending upon number of classes needed to be classified. Traditionally a sigmoid function is used for binary classification and a Softmax function is used for multi-class classification.

F. Ensemble Filtering with Convolutional Neural Networks

The application of individual filters on images did not yield appreciable accuracy compared to traditional CNN hence an ensemble of filters was developed to check the accuracy of the system. Here, ensemble of Watershed Filter, Adaptive Histogram Equalization and Adaptive Gaussian Threshold filters are used for preprocessing images prior to training on CNN.

III. Observations

A. Performance of different filters independently on CNN

Filters	**Accuracy**	
	Validation Accuracy	***Testing Accuracy***
Watershed Algorithm + CNN	88.68%	95.12%
Adaptive Gaussian Threshold + CNN	97.03%	100.00%
Adaptive Histogram Equalization + CNN	84.84%	97.56%
Gabor Filter + CNN	84.73%	54.78%
Sobel Filter + CNN	49.55%	0.00%
CNN	86.56%	98.78%

CNN parameters are Epochs = 4, Filter Layer = 32, Activation Layer =ReLu, Max Pooling Size = 2

B. Performance of Ensemble of Filters on CNN

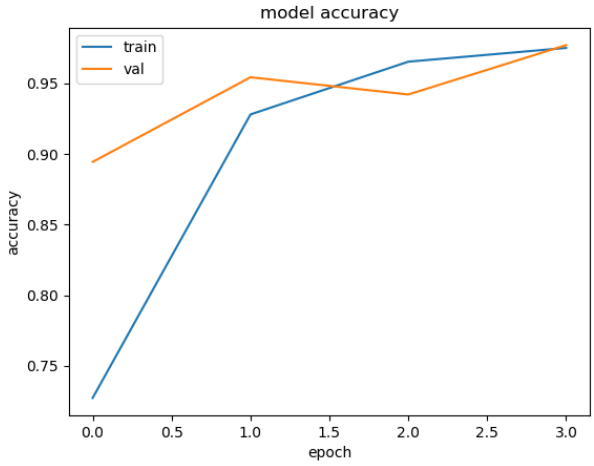


Fig1. Model accuracy with Ensemble Filters

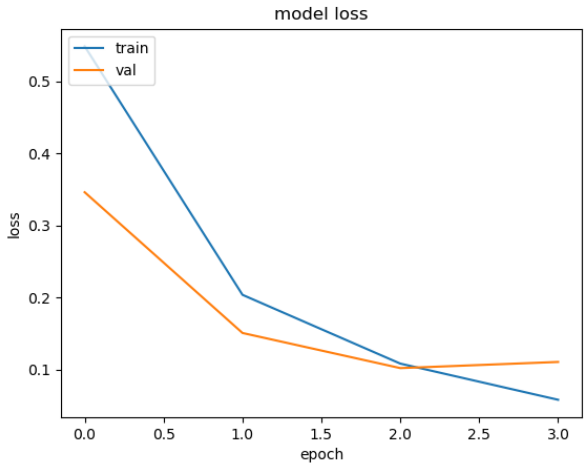


Fig 2. Model Loss with Ensemble Filters

Model with ensemble filters gave 99.20% accuracy as compared to traditional model

IV. Conclusions

The conclusion for this project is that we reach a null hythosis that the ensemble of filters with CNN work better in giving results than the normal CNN. The above mentioned disadvantages of the CNN has been handled by the ensemble of filters giving a better real time accuracy than expected. It is that the most of the filters that were individually trained and tested on the CNN were not that productive.

REFERENCES

[1] L. Pigou et al, Sign Language Recognition Using Convolutional Neural Networks. European Conference on Computer Vision 6-12, September 2014

[2] Garcia, Brandon and Viesca, Sigberto. Real-time American Sign Language Recognition with Convolutional Neural Networks. In Convolutional Neural Networks for Visual Recognition at Stanford University, 2016

[3] Sawant Pramada, Deshpande Saylee, Nale Pranita, Nerkar Samiksha, Mrs.Archana S. Vaidya: Intelligent Sign Language Recognition Using Image Processing, IOSR Journal of Engineering (IOSRJEN), Vol. 3, Issue 2, February 2013

[4] [Bauer & Hienz, 2000] Relevant feature for video- based continuous sign language recognition. Department of Technical Computer Science, Aachen University of Technology, Aachen, Germany, 2000.pages 440 – 445.

[5] Liu Yucheng and Liu Yubin, "Incremental Learning Method of Least Squares Support Vector Machine", International Conference on Intelligent Computation Technology and Automation" VCL-94-104, 2010

[6] T.Starner and A. Pentland. Real-Time American Sign Language Recognition from Video Using Hidden Markov Models. Computational Imaging and Vision, 9(1); 227-243, 1997.

[7] P. Mekala et al. Real-time Sign Language Recognition based on Neural Network Architecture. System Theory (SSST), 2011 IEEE 43rd Southeastern Symposium 14-16 March 2011.

[8] Richard Watson, "Gesture recognition techniques", Technical report, Trinity College, Department of Computer Science, Dublin, July, Technical Report No. TCD-CS-93-11, 1993