International Journal of Computer Trends and Technology
ISSN: 2231-2803 / https://doi.org/10.14445/22312803/IJCTT-V67I4P109

Volume 67 Issue 4, 40-43, April 2019
© 2019 Seventh Sense Research Group®

Original Article

An Improvised Word Recognition System by Hybridizing CNN and SIFT

Neethu Mohan[1], Arul V H[2]

[1,2] *Electronics and Communication Engineering, APJ Abdul Kalam Technological University*
Thejus Engineering College, Kerala, India

Abstract *- This paper mainly focuses on developing an efficient word recognition system by combining the good parameters of the SIFT and incorporating it with the CNN structure. Several advancements have been made in the Automatic Speech Recognition (ASR) technology that enables the machine to understand the natural language. The main constrain rise is the nature of the input speech signal, which makes it difficult to retain the original information. This can be overcome by hybridizing the (SIFT) Scale Invariant Feature Transform with (CNN) Convolutional Neural Network architecture. The noisy speech signal is initially passed through the pre-processing stage and converted to the spectrogram to extract the feature. The extracted features are now fed to the layers of CNN to train the model. The vectors are now cross-matched at the testing phase, and the maximum close weighted value from the fully connected layers leads to the output. The system performs with an efficiency of 94.78% in a non-isolated environment.*

Keywords *- ASR, SIFT, CNN, spectrogram*

I. INTRODUCTION

Studies on speech recognition and its processing have been carried out for more than five decades. The goal of an automatic speech recognition system is the transcription of human speech into spoken words. Speech recognition can be put as converting a speech signal to a sequence of words employing an algorithm implemented on a computer program[1].

Recently many sites have reported some successful results using deep convolutional neural networks instead of standard fully connected DNN's[9].

Speech or, more clearly, word recognition involves a pre-processing stage, segmentation, feature extraction, and classification. Pre-processing stage involves recognizing whether the signal is voiced or unvoiced and removing noise from the signal[1].

Feature extraction obtains different features such as power, pitch, and vocal tract configuration from speech signals. For speech recognition problems, some common methods are

Hidden Markov model, neural network, DNN, etc. Similarly, there are several feature extraction algorithms available for speech recognition.SIFT is such a feature extraction algorithm that is invariant to scale, translation and rotation and is the only feature extraction method found so far that can handle all three-time distortions (time scaling, time-stretching, pitch-shifting) simultaneously[8].

Machine learning is a category of algorithms that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. A neural network is a system of hardware or software patterned similar to the operation of neurons in the human brain. Neural networks are a variety of deep learning technology[2].

Neurons in a fully connected layer have full connections to all activations in the previous layer as seen in regular neural networks and work similarly [11]. Neural networks are adaptive, which means that they modify themselves as they learn from their initial training phase. CNN is a type of artificial neural network used in speech recognition and processing specifically designed to process pixel data. CNN uses powerful deep learning technology to perform both generative and descriptive tasks. CNN is mainly designed for reducing processing requirements.CNN merges two signal processing stages, i.e., feature extraction and classification, into a single one[4]. Further, the accuracy, speed of learning, classification, tolerance to irrelevant attributes, tolerance to redundant attributes, tolerance to noise, and attempts for incremental learning are efficiency-wise better for CNN than other supervised machine learning algorithms[5]. The layers of CNN consist of an input layer, output layer, and hidden layers[2].

The following sections are arranged: section 1 includes model architecture, section 2 includes results and discussions, and section 3 concludes the paper.

II. MODEL ARCHITECTURE

The model proposed here is an improvised architecture for word or speech recognition by hybridizing the advantages of both CNN and SIFT for pre-processing, feature extraction, and classification.

A. Hybridized CNN and SIFT

The database of the speech signal is created in a non-isolated noisy environment. In the pre-processing stage, the signal is made free from noise. This is done by filtering and windowing. The speech signal is then converted into its corresponding spectral components. Hence, the output is the spectrogram of each speech signal in the database.

The spectrogram of each signal or sound wave obtained is a digital image, a greyscale image with pixel values varying between 0 to 255. The spectrogram of each audio wave is then fed into SIFT for feature extraction. The output vectors of SIFT are then fed as input to CNN for further training and testing.

B. SIFT Architecture

SIFT is an algorithm used for extracting features from images in image recognition or image processing[7].SIFT uses Gaussian filtering for smoothing the original image[10]. Each signal recorded in a noisy environment is converted into its corresponding spectral components, i.e., its frequency domain. Hence the spectrogram of the entire database is generated. This spectrogram is generated since SIFT algorithm extracts distinctive invariant features from digital images in a way invariant to scale, translation and rotation. Let the spectrogram of each audio signal be denoted as I(x,y).

The spectrogram is smoothened by noise removal before the images are provided to SIFT for feature extraction. Then the noise removed spectrogram of each audio signal was given to SIFT for feature extraction. The image of each spectrogram is first used to generate the scale space for SIFT. After that, the LOG approximation is done to find key points in the image.

The next step is to compute the local maxima or minima and find the subpixel maxima or minima. The key points with contrast less than a certain pixel value are rejected to get rid of low contrast features.
Next, SIFT calculates the keypoint orientations. The last step includes generating features. These extracted SIFT features are stored as digital images, say L(x,y). Hence, the entire database for all the audio signals is created. These features generated are stored as greyscale images, and the features extracted are pixel values that differ according to the keypoint extraction. Figure 1 shows the SIFT algorithm.

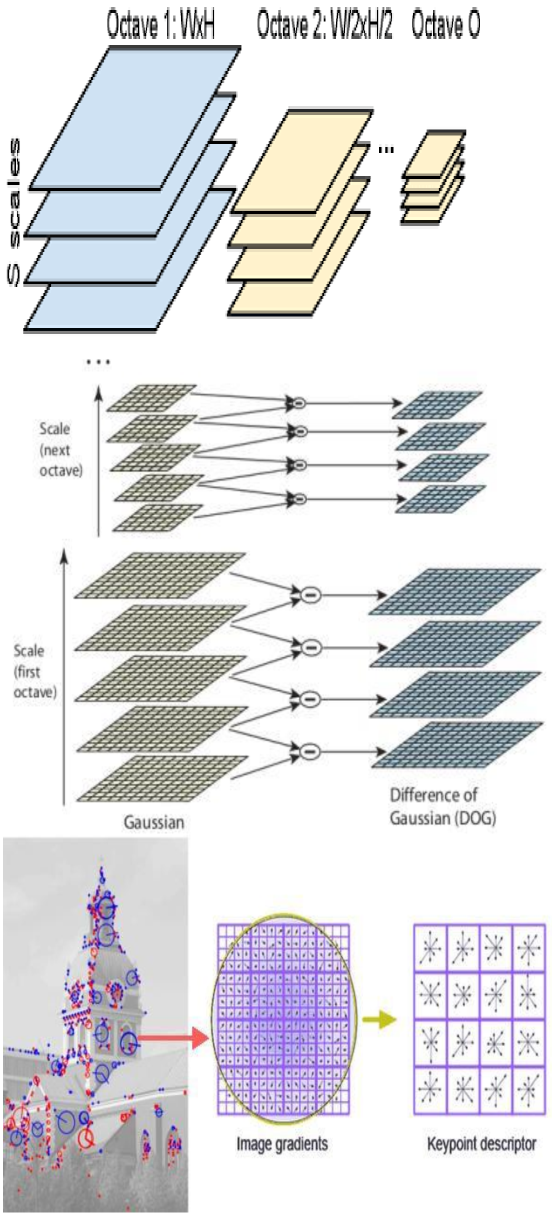

Fig. 1 Feature extraction using SIFT

C. Convolutional Neural Network

The output of SIFT algorithm (L(x,y)) is provided to CNN (Convolutional Neural Network) for further classification. CNN provides a better method of using the information in the training set to build multiple layers of non-linear feature detectors[6]. CNN merges the two feature extraction and classification stages, reducing the processing requirements[4]. But we still use SIFT for feature extraction because algorithms like CNN require feature maps of the same size as inputs. Else it would require an initial stage of resampling and quantization which reduces the algorithm's flexibility in large-scale speech perception. This issue is solved in SIFT because SIFT can handle feature maps of different sizes. Hence the initial resampling and quantization required in CNN can be avoided here. Hence SIFT is more flexible since it is invariant to scale, rotation, and translation[3]. The first need is to prepare the CNN layers for classification using CNN.

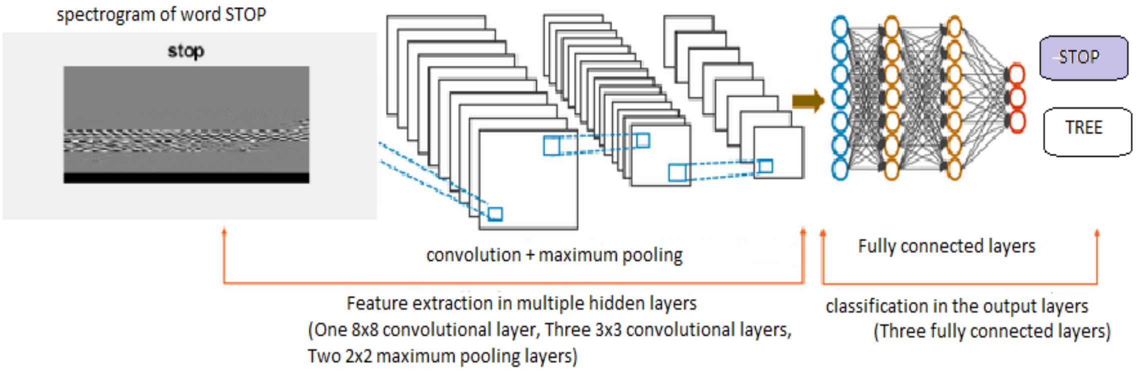
spectrogram of word STOP
stop
convolution + maximum pooling
Fully connected layers
STOP
TREE
Feature extraction in multiple hidden layers
(One 8x8 convolutional layer, Three 3x3 convolutional layers,
Two 2x2 maximum pooling layers)
classification in the output layers
(Three fully connected layers)

Fig. 2 CNN Architecture

Suppose K is the total number of feature maps. Each hidden layer of CNN has several units, each of which takes all outputs of the lower layer as input, multiplies them by a weight vector, sums the result, and passes it through a non-linear activation function such as sigmoid or tanh as given below[1]:

$$O_i^{(l)} = \sigma\left(\sum_j O_j^{(l-1)} w_{j,i}^{(l)} + w_{o,i}^{(l)}\right) \quad (1)$$

$O_i^{(l)}$denotes the output of the i-th unit in the *l*th layer. $w_{j,i}^{(l)}$denotes connecting weights from the j-th unit in layer *l*-1 to the i-th unit in the *l*th layer. $w_{o,i}^{(l)}$is a bias added to the i-th unit and σ (x) is a non-linear activation function[1].

$O_i$for (*i* = 1,2,3……K), is connected to many feature maps; say J number of feature maps. O_i is also connected to $w_{i,j}$ (*i* = 1,2,3….K and *j* = 1,2,3….J). Each unit of one feature map in the convolution layer can be computed as;

$$q_{j,m} = \sigma\left(\sum_{i=1}^{K}\sum_{n=1}^{F} O_{i,n+m-1} w_{i,j,n} + w_{o,j}\right) \quad (2)$$

(j = 1,2,3……J)[1].

The purpose of the pooling layer is to reduce the resolution of feature maps. The pooling function is independently applied to each convolution feature map [1]. In the proposed system, maximum pooling is done. The max-pooling is done as

$$P_{j,m} = max_{n=1}^{G} q_{i,(m-1)*s+n} \quad (3)$$

Here G is the pooling size, and S is the stride size(overlapping adjacent pooling windows)[1].

The approach proposed four convolutional layers, two pooling layers, and three fully-connected layers. Are constructed. Figure 2 represents the CNN architecture proposed in this paper.

D. Database

The database consists of sound waves of different pitches and frequencies of people aged 25 to 40 who speak out the words "stop" and "tree ."This database is taken from Tensor Dataflow. Seven hundred samples of word tree and 500 samples of word stop are taken for testing and training. Each sound wave is converted to its corresponding spectrogram image and stored as a greyscale image in the database.

III. RESULTS AND DISCUSSIONS

The output of SIFT has greyscaled images with pixel values ranging between 0-255, each of which is processed out as 8x8 vector blocks. These feature maps are again stored as greyscale images of size 200x88. This is then given to the convolutional layer of CNN, having a convolving filter of size 8x8 and 16 layers. This is to capture major changes in images. Further, 3 more convolutional layers are patterned, each of which has a 3x3 filter with 32 layers, 64 layers, and 128 layers. Between the four convolutional layers, two pooling layers of size 2x2 are patterned for maximum pooling. Three fully connected layers (200,100, and 2, respectively) are designed to provide information about convolutional layers and pooling layers. The third fully connected layer gives the classified output. The iteration accuracy is 94.78%, i.e., the hybridized system works with an overall accuracy of 94.78%.

Table 2 shows the classification accuracy of different recognition modules in speech recognition using CNN alone. Hence, the result shows that the accuracy is not more than 83.9% for different recognition modules. The system proposed in this paper gives a classification accuracy of 94.8%.

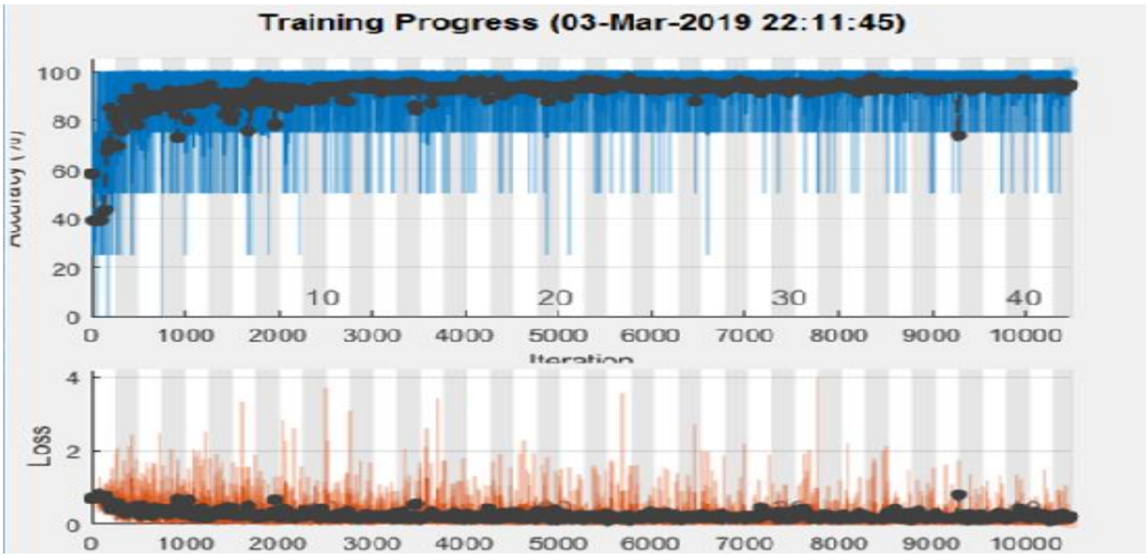

Fig. 3 Training progress

TABLE I:Iteration results

Validation Accuracy	94.78%
Training finish mode	Manual
Training cycle epoch	42 of 1000
Iterations per epoch	251
Iterations	10480 of 251000
Maximum iterations	251000
Validation frequency	20 iterations
Time elapsed till the present iteration	527 minutes 55 seconds

TABLE II:Classification accuracy of different recognition modules[12]

Recognition Module	C180	C360	C720
Classification Accuracy	82.2%	83.4%	83.9%

IV. CONCLUSION

Vast advances have been made in ASR technology, and its crucial objective is to make the machine understand natural language. SIFT is a better feature extraction algorithm that handles all the time-frequency discriminations all at one time.

CNN is a deep learning technology that reduces the pre-processing requirements by merging two stages of signal processing, i.e feature extraction and classification, together. But CNN requires feature maps of the same size, which requires an initial resampling and quantization. This discrepancy can be removed while using SIFT. The system proposed in this paper hybridizes these advantages of both SIFT and CNN and iterates with a validation accuracy of 94.78%.

REFERENCES

[1] Ossama Abdel-Hamid, Abdel-rahman Mohamed, Hui Jiang, Li Deng, Gerald Penn, and Dong Yu, "Convolutional Neural Networks for Speech Recognition," IEEE/ACM transactions on audio, speech, and language processing, vol. 22, no. 10, October 2014.

[2] https://searchenterpriseai.techtarget.com

[3] Quang Trung Nguyen, The Duy Bui, "Speech classification using SIFT features on spectrogram images," Vietnam Journal of Computer Science, 3(4), 247-257.

[4] https://www.quora.com/How-is-convolutional-neural-network-algorithm-better-as-compared-to-other-image classification-algorithms.

[5] Osisanwo F.Y, Akinsola J.E.T, Awodele O, Hinmikaiye J.O, Olakanmi O, Akinjobi J, "Supervised Machine Learning Algorithms: Classification and Comparison," International Journal of Computer Trends and Technology (IJCTT) – Volume 48 Number 3 June 2017.

[6] Geoffrey Hinton, Li Deng, Dong Yu, George Dahl, Abdel-rahman Mohamed, NavdeepJaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Brian Kingsbury, Tara Sainath, "Deep Neural Networks for Acoustic Modeling in Speech Recognition**," IEEE Signal Processing Magazine** | November 2012, Vol 29: pp. 82-97.

[7] ishack.in/tutorials/sift-scale-invariant-feature-transform-introduction

[8] Xiu Zhang, Bilei Zhu, Linwei Li, Wei Li, Xiaoqiang Li, Wei Wang, Peizhong Lu, and Wenqiang Zhang, "SIFT-based local spectrogram image descriptor: a novel feature for robust music identification," Zhang et al. EURASIP Journal on Audio, Speech, and Music Processing (2015) 2015:6.

[9] Jui-Ting Huang, Jinyu Li, and Yifan Gong, "An Analysis of Convolutional Neural Networks for Speech Recognition," Microsoft Corporation, One Microsoft Way, Redmond, WA 98052.

[10] Tomoaki Yamazaki, Tetsuya Fujikawa, Jiro Katto, "Improving the performance of SIFT using bilateral filter and its application to generic object recognition," 2012 IEEEInernational Conference on Acoustics, Speech, and Signal Processing, ICASSP 2012- Kyoto

[11] Harsh Pokarana, "Explanation of Convolutional Neural Network," IIT Kanpur.

[12] Tao Wang, David J. Wu, Adam Coates Andrew Y. Ng; "End-to-End Text Recognition with Convolutional Neural Networks"; Stanford University, 353 Serra Mall, Stanford, CA 94305