CENG476 INTRODUCTION TO MACHINE LEARNING
LIP READING

161180030 Demet EROL - 161180067 Müzeyyennur YILGIN
RAICHU

Introduction

The World Health Organization estimates that there are over 460 million deaf people in the world. These people need sign language to communicate in society. However, because sign language is not widely used, communication is difficult. To address this issue, various solutions such as hearing aids and speech-to-text translation applications have been developed.Since external sounds have a negative impact on speech perception in these systems, adequate reliability cannot be attained in real life. Lip reading, on the other hand, uses visual data to eliminate noise-induced issues [1].

Lip reading is the identification of spoken words as well as the study of lip formation. Lip reading has become one of the artificial intelligence experiments as technology has advanced. [2] Areas of use include Aviation, military systems, defense, healthcare, sports refereeing, and the translation of historical documentaries and films [3].

In the project MIRACL-VC1 dataset [4] is used. It containing both depth and color images of fifteen speakers uttering ten words and ten phrases, ten times each. The sequence of images represents low quality video frames. The data set contains 3000 sequences of varying lengths of images of 640 x 480 pixels, in both color and depth representations, collected at 15 frames per second. The lengths of these sequences range from 4 to 20 image frames. The words are listed below.

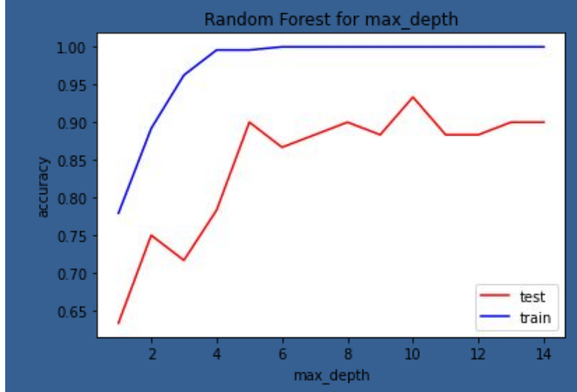
ID	WORDS
1	Begin
2	Choose
3	Connection
4	Navigation
5	Next
6	Previous
7	Start
8	Stop
9	Hello
10	Web

The large amount of data caused the preprocessing step and training to take a long time. For this reason, the words "begin", "choose", "connection"; 5 women and 5 men were selected as speakers. The Dlib library was used to detect the faces in the images. The lip is represented by points between 49 and 68 in the “shape predictor 68 face landmarks.dat” model. Lip images were created using this model, resized, and stored in a separate folder.

With the pixels from the cut lip images, speaker ids, and targets, a dataframe was created. The lips were cut from the images and recorded. After the resize process, the padding process was applied for the data with less than 20 data. Then the data were normalized with the min max scaler.

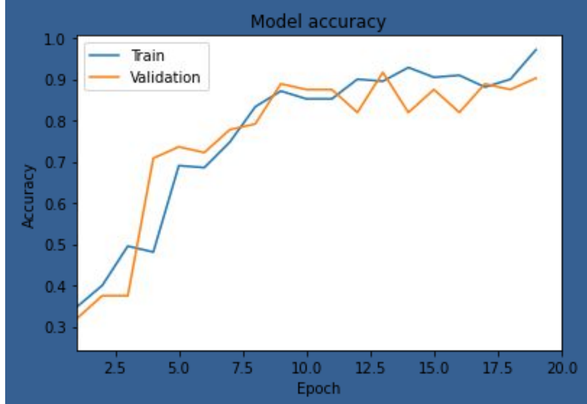
Machine Learning Algorithms

A dataframe containing the pixel values of the lip images, the speaker and the word was created. The data was split into two parts: 80 percent train and 20 percent test. Since they are excellent at classification, models were generated as the KNN, Decision Tree, SVM, Random Forest. After the test with different parameter values, the most successful algorithm was Random Forest.



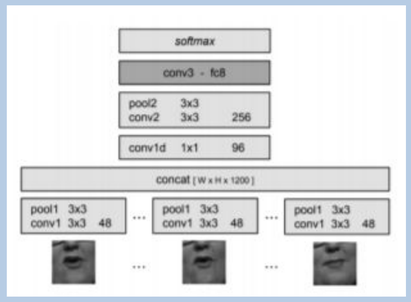
CNN

After the scale operation, y values were converted to categorical data. 0th index "begin", 1st index "choose", 2nd index "connection". In order to adapt the data to the model, the dimension was determined as 5. The data is divided into 70% train, 30% test (80% validation, 20% test). Conv3D and CNN model consists of 3 layers. Pool size as (2,2,2) and Relu as the activation function is selected. While calculating the loss after the dense and dropout steps, categorical_crossentropy was preferred and the optimizer was "Adam". After the model is trained with batch size 21, epoch 20, the test results are as follows.

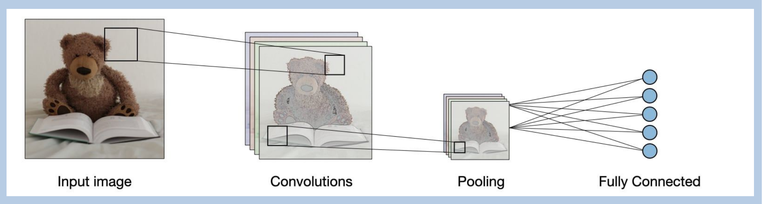


Convolutional Neural Networks

CNN is a deep learning algorithm consisting of different trainable layers. These layers are Convolution Layer, Non-linearity Layer, Pooling Layer and single or multiple fully connected layers. The Convolutional layer consists of many learnable filters. The feature map is created by hovering the filters over the entire image. By changing the filter coefficients, important regions on the image are determined [5].

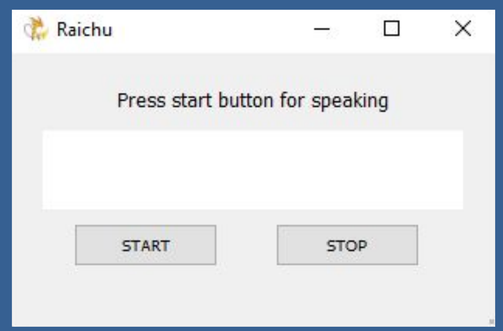


Chung J.S., Zisserman A. (2017) Lip Reading in the Wild. In: Lai SH., Lepetit V., Nishino K., Sato Y. (eds) Computer Vision – ACCV 2016.



https://stanford.edu/~shervine/teaching/cs-230/cheatsheet-convolutional-neural-networks

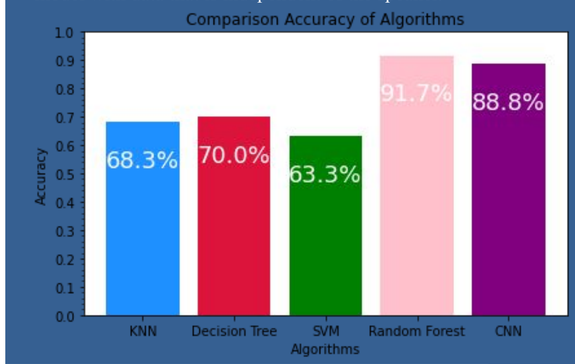
Interface



The created model was saved with the save function of the Keras library. The interface seen in Figure 3.11 was designed with Pycharm and Qt Designer. Lips were detected with “shape predictor 68 face landmarks.dat” on video images taken from the user using OpenCV. After pressing the start button, the video recording was started and the user was expected to say one of the words "begin", "choose" and "connection". After pressing the stop button, one of every 5 frames was taken from the video recording and preprocessing steps such as resize and scale were performed. After the test data created was given to the model, the prediction made was transferred to the user as text with QLabel.

Conclusion

Lip reading has been a significant need in a variety of fields today. In this application created with the words "begin", "choose", and "connection" from the Miracl-VC1 dataset. CNN, KNN, SVM, Decision Tree, and Random Forest were used. Random Forest, with 91.7 percent accuracy, and CNN, with 88.8 percent accuracy, are the highest. In the next step was created a gui to validate the CNN model with data that is independent of the speaker.



References

[1] Abrar, M.A., Islam, A. N. M. N., Hassan, M. M., Islam, M. T., Shahnaz C. and Fattah, SA., "Deep Lip Reading-A Deep Learning Based Lip-Reading Software for the Hearing Impaired", *2019 IEEE R10 Humanitarian Technology Conference*, 2019, 40-44.

[2] Khan, S., Azmi, H & Nair, A. and Mirza, H., “Implication and Utilization of various Lip Reading Techniques”, *International Journal of Computer Applications*, 2017, 167, 25-27.

[3] Burton, J., Frank, D., Saleh, M., Navab, N. and Bear, H. L., "The speaker-independent lipreading play-off; a survey of lipreading machines", *2018 IEEE International Conference on Image Processing, Applications and Systems (IPAS)*, 2018, 125-130.

[4] https://sites.google.com/site/achrafbenhamadou/-datasets/miracl-vc1

[5] Fattah, P., Salihi, N. K., Rashid, T. A. and Shamsaldin, A. S. (2019), The Study of The Convolutional Neural Networks Applications, UKH Journal of Science and Engineering, 31-40.