Chinese Sign Language Classification: Literature Review

Problem Formulation:

Nowadays, about 5% of the population suffers from the issue of hearing loss in the world. There is also a large number of hearing-impaired people in China, number of which 27.9 million. [1] For people who have trouble hearing, sign language is the easiest form of communication. Sign language recognition is an important and difficult task when it comes to converting sign language into text. The current sign language recognition methods are broadly divided into two types: the traditional machine learning method using image features and the deep neural network-based method. The former uses machine learning techniques to perform classification based on the image features of the interested area. The former uses traditional image segmentation algorithms to segment hand shapes from sign language images or video frames of sign language video.

Literature Review:

[2]An efficient technique for the recognition of Indian Sign Language ISL letters, words, and numbers used in daily life is provided in this study. Convolutional layers are the first in the proposed CNN architecture, followed by ReLU and max-pooling layers. Different filtering window sizes are included in each convolutional layer, which helps to increase recognition speed and precision. A web camera-based dataset of 35,000 images from 100 static signs has been generated. The proposed architecture has been tested on approximately 50 deep-learning models using different optimizers. The system results in the highest training and validation accuracy of 99.17% and 98.80%, respectively, with respect to different parameters such as the number of layers and filters. A variety of optimizers were used to test the suggested system, and it was discovered that SGD beat Adam and RMSProp optimizers, with training and validation accuracy on the grayscale picture dataset of 99.90% and 98.70%, respectively. The proposed system is robust enough to learn 100 different static manual signs with lower error rates. It has been found that the system outperformed other existing systems even with a smaller number of epochs. The major source of challenge in sign language recognition is the capability of sign recognition systems to adequately process a large number of different manual signs with low error rates.

[3]To classify selfie sign language gestures, they suggested using the CNN architecture. Four convolutional layers make up the CNN architecture. The consideration of each convolutional layer with a particular filtering window size increases identification speed and precision. The benefits of the max and mean pooling techniques are combined in a stochastic pooling strategy. A total of 300000 sign video frames were produced when we generated the selfie sign language data set using 200 ISL signs with 5 signers in 5 user-dependent viewing angles for 2 seconds each at 30 frames per second. To determine the reliability of the large training modes needed for CNNs, training is done in many batches. The training is carried out with three sets of data (i.e., 180000 video frames) in Batch-III in order to maximize the recognition of the SLR. This CNN architecture has higher training accuracy and validation accuracy than the previously suggested SLR models which were Mahalanobis distance classifier (MDC), Adaboost classifier and artificial neural network

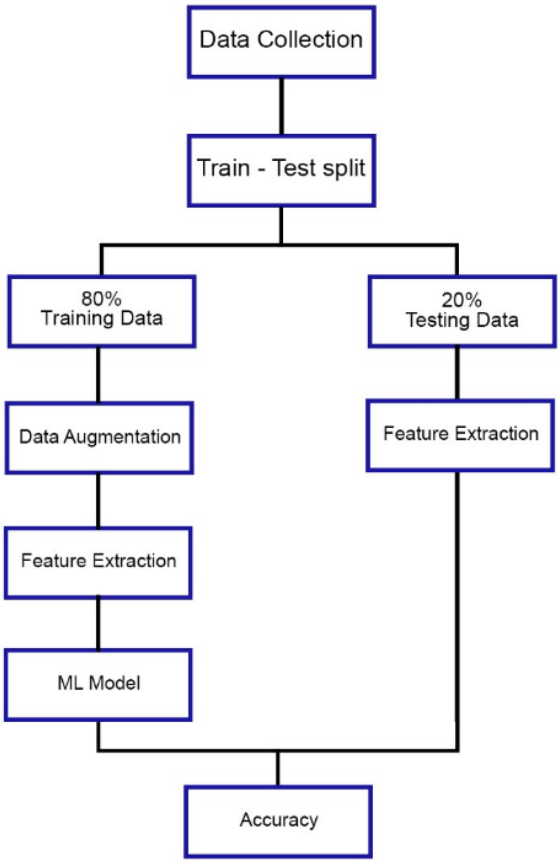
(ANN). The suggested CNN architecture shows less loss during training and validation. Comparing the proposed CNN model to existing state-of-the-art classifiers, the recognition accuracy rate is higher at 92.88%.

[4]The system can detect one or two hands in a video stream in real-time. The pipeline of the experiment is as follows. Capturing frames using a Logitech HD C310 webcam. Then he used the Open-CV library to find hand contour, convex hull, and defect points, palm center localization and stabilization, fingertips identification, and finally used SVM as the classifier. After identifying individual fingertips, gestures can be classified by detecting the number of fingers. If it is five, then it is an open palm and if it is zero, then it is a closed palm. To do this, a Support Vector Machine classifier was used which acts as a separating hyperplane. Regarding the results, the best accuracy achieved was over 85%. The model worked well for every gesture. However, it is not ideal, because not all the fingers are classified correctly for every case.

[5]The authors used feature reduction by applying LDA or PCA then tried different supervised machine learning algorithms to classify Electromyography data into 7 different human hand gestures. The traditional classification algorithms were K- Nearest Neighbor (KNN), Naive Bayes, Decision Tree, and Random Forest. They used also deep learning classifiers such as Artificial Neural Networks (ANN) and Long Short-Term Memory (LSTM). For the results, the Random Forest classifier is the best model that achieves 99.43% accuracy and the LSTM model gives 99.19% accuracy. The models are reliable and applicable because they calculated the accuracy and error graph for the train, test, and validation set. EMG data patterns for different hand gestures have differences from each other, so it is easy for the Random Forest ensemble to classify them perfectly.

Proposed Solution:

- We have taken 500 images of numbers with Chinese hand signs for classification.
- Split the images into training and testing 80:20 respectively.
- As the number of training data samples is small we will apply different data augmentation techniques (Flipping, Translation, Cropping, etc.).
- Using transfer learning to fine-tune the pre-trained AlexNet [6] model on our images and compute the accuracy of the test data to represent our baseline.
- Apply 2 different feature extraction techniques, the first technique is by using The Zernike [7] moment after applying a binary mask on the images for hand pixels to be 1 and 0 otherwise. We know that the convolution layers act as feature extractors so we collect the features from the FC6 layer [8] of the fine-tuned AlexNet model.
- Combining the extracted features from both of the techniques and feeding them into different ML models for example LIBSVM [9], Multi-layer Perceptron (MLP), Random Forest, and XGBoost. Then, select the champion model based on the test accuracy, and compare the results with the baseline.



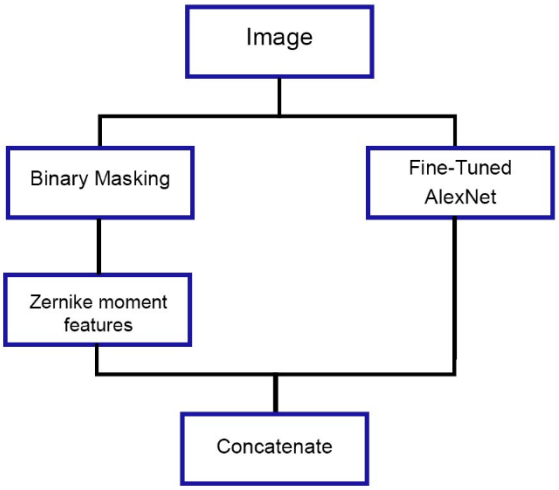


Figure 1: Feature Extraction

Figure 2: Project Pipeline

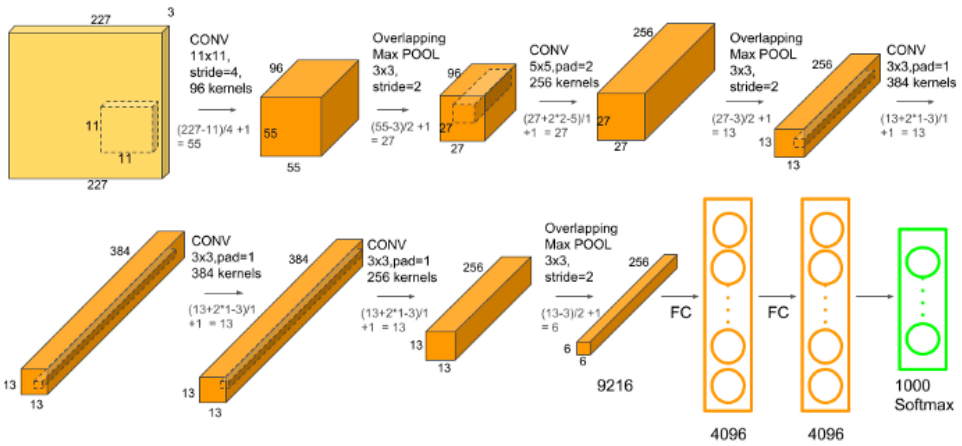


Figure 3: AlexNet Architecture[10]

References:

[1] X. Jiang, B. Hu, S. Chandra Satapathy, S. H. Wang, and Y. D. Zhang, "Fingerspelling Identification for Chinese Sign Language via AlexNet-Based Transfer Learning and Adam Optimizer," *Sci Program*, vol. 2020, 2020, doi: 10.1155/2020/3291426.

[2] A. Wadhawan and P. Kumar, "Deep learning-based sign language recognition system for static signs," *Neural Comput Appl*, vol. 32, no. 12, pp. 7957–7968, Jun. 2020, doi: 10.1007/s00521-019-04691-y.

[3] G. A. Rao, K. Syamala, P. V. V. Kishore, and A. S. C. S. Sastry, "Deep convolutional neural networks for sign language recognition," in *2018 Conference on Signal Processing And Communication Engineering Systems, SPACES 2018*, Mar. 2018, vol. 2018-January, pp. 194–197. doi: 10.1109/SPACES.2018.8316344.

[4] G. N. Pham, "International Journal of Multidisciplinary Research and Growth Evaluation Experiment hand gesture recognition and classification using machine learning algorithm," vol. 2, no. 5, pp. 337–339, [Online]. Available: www.allmultidisciplinaryjournal.com

[5] A. B. Habib, F. bin Ashraf, and A. Shakil, "Finding Efficient Machine Learning Model for Hand Gesture Classification Using EMG Data," in *2021 5th International Conference on Electrical Engineering and Information and Communication Technology, ICEEICT 2021*, 2021. doi: 10.1109/ICEEICT53905.2021.9667856.

[6] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks." [Online]. Available: http://code.google.com/p/cuda-convnet/

[7] Zernike, F.: Beugungstheorie des Schneidenverfahrens und seiner verbesserten Form, der Phasenkontrastmethode. Physica **1**(7–12), 689–704 (1934)

[8] Barbhuiya, A.A., Karsh, R.K. & Jain, R. A convolutional neural network and classical moments-based feature fusion model for gesture recognition. *Multimedia Systems* **28**, 1779–1792 (2022). https://doi-org.proxy.bib.uottawa.ca/10.1007/s00530-022-00951-5

[9] Chang, C.-C., Lin, C.-J.: LIBSVM: a library for support vector machines. ACM Trans Intell Syst Technol (TIST) **2**(3), 1–27 (2011)

[10] https://medium.com/analytics-vidhya/concept-of-alexnet-convolutional-neural-network-6e73b4f9ee30