COMP9444 Project Summary

A new deep learning neural network structure to recognize static gestures in sign language.

<Xinjun Tan, z5432003; Lingyun Xiao, z5403246; Yang Song, z5441033; Hongzhu Wang z5427768; Hao Zeng, z5445039 >

Group: import genshin

I. Introduction

The objective of this project is to reproduce the model from the paper "Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter".

In this work, a new deep learning neural network structure for the recognition of static hand gestures in sign language design and implement. The new structure combined CNN and classical non-intelligent feature extraction methods, ORB descriptor and Gabor filter.

The importance of this work lies in the broad applications of gesture recognition in various fields such as human-computer interaction, assisting people with disabilities, and virtual reality. However, due to the complexity and variability of gestures, static gesture recognition still faces challenges in uncertainties such as rotation and ambiguity of gestures.

The solution proposed in this project is a novel structure that combines Convolutional Neural Network (CNN) and classical non-intelligent feature extraction methods, ORB discriptor and Gabor filter. In this structure, the hand gesture image, after preprocessing and background removal, passes through three different streams of feature extraction, each independently extracting specific features, to effectively extract effective features and determine the class of the hand gesture. These three streams consist of three methods widely used in hand gesture classification: CNN, Gabor filter, and ORB feature descriptor. Then these features are merged and form the final feature vector.

By combining these efficient methods, we not only achieve a very high accuracy in hand gesture classification, but also make the proposed structure more resistant to uncertainties such as rotation and ambiguity in the hand gestures. Another outstanding feature is that compared to similar methods, our structure can be widely applied to different image databases. The transfer learning technique demonstrates that our structure can be used as a pre-trained structure for any type of image database.

In the final experiment, we applied the proposed structure to the ASL Alphabet dataset. The results show that for the 26,100 test images of the ASL Alphabet, our structure achieved a mean accuracy of 99.80%, demonstrating the effectiveness of our method.

Data Source:

The ASL Alphabet Database is a collection of images used for machine learning and AI recognition tasks, specifically aimed at recognizing American Sign Language (ASL) hand gestures.

(https://www.kaggle.com/datasets/grassknoted/asl-alphabet)

Each image in the database represents one of the 26 letters in the English alphabet as signed in ASL. The images are typically captured against a variety of backgrounds to simulate real-world conditions, making it an ideal resource for training robust recognition models.

The database is an important tool in the field of AI, particularly for projects aimed at improving accessibility technology for the deaf and hard-of-hearing communities. Its use allows researchers and developers to test and validate the accuracy and effectiveness of their models in recognizing and interpreting ASL.

II. Methods

In our project, we used CNN, Gabor filter, ORB feature descriptor, K-means and Canny edge detector.

1. Canny method: Canny edge detector method is applied to detect sharp discontinuities or the edges of hand gesture in the 200 × 200 input image which can reduce the noise of image background and can be easily segmented from the input image.

2. ORB method: ORB descriptor is a combination of FAST algorithm and BRIEF algorithm, which can enhance robustness of structure to rotation, scaling, and light intensity variations.

3. Gabor filter: Gabor filter can extract texture information in different directions.

4. K-means: K-means is an unsupervised machine learning algorithm primarily used for clustering analysis, which can partition a set of observations into a number of clusters (K), each observation belonging to the cluster with the nearest mean.

5. CNN: We have two CNNs with the same architecture, but one for processing images processed by Gabor filter, the other for directly processing the original image. In this way, from two different perspectives, two CNNs are used to exploit all the useful features in identifying the hand gesture class from the input image. Also the dropout technique is used to prevent over-fitting and to make the robust structure.

The accuracy of intelligent methods are higher than the classical methods, and on the other hand, the classical methods have the more effective extracted features in the field of hand gesture recognition. Thus, we design a new deep learning neural network including the CNN and classical non-intelligent feature extraction method to identify the hand gesture in the sign language. In addition, the reason to select the ORB feature descriptor in comparison to the SURF and SIFT feature descriptors is that the ORB is at two orders of magnitude faster that these descriptors.

III. Experimental Setup

Our experiments are based on the open-source dataset obtained from Kaggle. The dataset consists of 29 classes, representing hand sign alphabets from A to Z, along with three special signs for nothing, space, and delete. There are a total of 87,000 images in the dataset, with each image having a size of 200x200 pixels. The dataset is split into three subsets for training, validation, and testing. The training set comprises 60% of the data, amounting to 52,200 images, while the validation set contains 10% (8,700 images), and the test set holds 30% (26,100 images).

Data Preprocessing: We have four data preprocessing steps.

1.Resize the input image to 200x200 and convert it to a grayscale image, followed by converting it to an 8-bit unsigned integer.

2. Extract features using the ORB feature descriptor, apply Canny edge detection on the image, and obtain feature values and vectors.

3. Utilize the k-means clustering algorithm to cluster the feature vectors, resulting in 150 clusters' mapping, which will serve as input features for the neural network.

4. Resize and normalize both the original image and the image processed with Gabor filters, which will act as inputs for the neural network.

Evaluation Strategy: For evaluation, a comprehensive strategy was adopted. The test set was divided into four data folders: *test_noise*, *test_rotate*, *test_downsam*, and *test_origin*. These folders facilitated evaluating the model's performance under various conditions, considering different shooting conditions, rotations, downsampling, and the original data.

To reduce training randomness' impact, we conducted multiple runs and averaged the results for more reliable evaluations. This approach helped understand the model's generalization and performance across different subsets of the test data.

Key Model Hyperparameters: During model training, key hyperparameters were optimized to enhance performance. These included learning rate, batch size, number of epochs, and regularization strength. Experimentation and validation were conducted to find the right balance between model performance and computational efficiency.

Challenging aspects:

Large data may require efficient methods and hardware for faster training.
Sign language variations may need complex models and additional data for generalization.
Diverse shooting conditions can lead to unstable predictions.
ORB feature extraction and Gabor filters may impact model performance.
Random selection can cause class imbalance, affecting recognition.
Comprehensive evaluation needed for testing dataset's variations.

IV. Results

	noise	rotation	Blurring	Ground true
MobileNetV2	8.48%	44.41%	80.76%	96.83%
ResNet50	11.17%	44.55%	82.07%	92.83%
LeNet	89%	26.62%	92.62%	92.13%
VGG16	53.79%	33.72%	99.58%	99.93%
Ours	67.66%	44.69%	96.97%	99.79%

(Comparison on The ASL Alphabet Database)

Main Findings: Our model shows significant robustness and adaptability to various conditions. Particularly, it performs at par with the best-performing models in each of the tested conditions. Its weakest area is noise handling, but even there it outperforms most other models.

Our model demonstrated robust performance across all tested conditions (noise, rotation, blurring). It achieved the highest accuracy in handling rotation conditions among all compared models (44.69%). For noisy conditions, our model attained 67.66% accuracy, performing second-best after LeNet. In blurring conditions, our model reached 96.97% accuracy, surpassed only slightly by VGG16. The robustness to blurring and noise can be attributed to our network architecture, which effectively reduces noise and background interference during the preprocessing stage.

V. Conclusions

The integrated model we reproduced from the paper performs relatively well in a variety of scenarios where the dataset is disturbed.

Although it is not the most outstanding performance in all scenarios, at the same time, it also has no particularly obvious shortcomings compared to the other models, and has been maintained in a comparatively good state, which indicates that the model has a good generalization in various scenarios of the gesture recognition task. It is hypothesized that this may be due to the complementary effects brought about by the ensemble approach.

Strengths: The model performs well in a variety of scenarios, is adaptable to various transformations (e.g., noise and blurring), and it also performs well on the original image, providing good generalizability.

Limitations: When a single disturbance of some kind is added to the test set, there always exists some pre-trained model that performs better than the one we reproduce. Therefore, the model we reproduce is not the best choice when the main type of disturbance is known.

Possible future work:

Ensemble methods: Continuing to combine with other models with resistance to specific disturbances may be able to further fill the disadvantage in the corresponding scenarios.

Fine-tuning and Transfer learning: Applying transfer learning and fine-tuning with pre-trained models may be able to improve the performance of our model in certain scenarios.

Dataset expansion: Expanding the training dataset with more diverse and representative data may lead to more robust and accurate models.

Reference:

Mahin Moghbeli Damaneh, Farahnaz Mohanna, Pouria Jafari. Static hand gesture recognition in sign language based on convolutional neural network with feature extraction method using ORB descriptor and Gabor filter. Expert Systems with Applications, Volume 211, 2023,118559, ISSN 0957-4174, https://doi.org/10.1016/j.eswa.2022.118559