IN3060/INM460 Computer Vision Coursework report

- **Student name, ID and cohort:** Vighnesh Mane (210021607) - PG
- **Google Drive folder:** https://drive.google.com/drive/folders/1zqbKvF3Y6VsP66FMna1vbP5KVnCmUUJb?usp=sharing

Data

The dataset utilised for this project is an image data collection comprised of diverse facial images displaying various facial image moods. The source of this dataset is the RAF database. Basic emotion labels include startled, scared, disgusted, happy, sad, angry, sadly angry, tragically fearful, and many others; however, for the purposes of this project, there are seven labels that are emphasised: surprised, fearful, disgusted, happy, sad, angry, and neutral. This data set consists of the training set and the testing set. For training, eighty percent is assigned, while twenty percent is designated for testing. The training set consists of 12271 aligned facial photos with basic expressions, whereas the testing set consists of 3068 aligned facial images with basic expressions.

In addition, emotion labels are separated into a training set and a testing set. In addition to this data set, another video data set is explored that will aid in facial expression recognition. This information was acquired from YouTube and is associated with the office series. This experiment identifies the facial emotions of three to four individuals with diverse face sizes and shapes using computer vision algorithms.

Implemented methods

This study implements the SVM-SIFT, SVM-HOG, MLP-SIFT, MLP-HOG, and CNN models. The initial model to apply was SVM-SIFT. The data used to build this model included 12271 train data and 3068 test data. The SIFT method has a high tolerance for scaling, rotation, brightness fluctuations, and noise, to begin with. SVM may get superior results with a minimal training set sample size and has high generalisation capability [1]. Initially, its fundamental model was developed. Consequently, the base model was trained with 12271 training data, and its performance was subpar; nevertheless, after adding the hyper parameters, the model's performance decreased marginally.

The second model developed was called SVM-HOG. Again, the same data was utilised for this model, consisting of 80% for the train side and 20% for the test side. First, the base model was trained, and its performance was superior than that of the SVM-SIFT model. Human detection utilising. HOG-SVM in hardware has a higher classification rate and throughput than deep learning algorithms [2]. After applying grid search to this model, its performance remained superior to that of the SVM-SIFT model.

The MLP-SIFT model was the third implemented model. This model was developed and trained in a manner similar to the previous two. First, the basic model was developed, whose performance was inferior to that of the second model but comparable to that of the first model. The system is straightforward, more accurate, and less complicated than other comparable systems [3]; After applying grid search to the model, its performance was comparable to that of the first model but diminished in comparison to the second.

The fourth model implemented was the multilayer perceptron with the hog. In this implementation firstly the base model was trained on eighty percent of the data. The performance of this model increased as compared to the third model. After applying grid search to this model, the performance of this model further even increased and became better as compared to the third model. Hog is used to divide the image into small regions called cells, and for each cell, compute a histogram of gradient directions or edge orientations. Discretize each cell into angular bins according to the gradient orientations. Group adjacent cells which are considered as spatial regions called blocks [4].

The last model was the CNN model and this model was the outperformed the rest of the models. For this model the implementation technique was same as compare to the other models. For creating this model data pre-processing was done and the data was modified little bit like a new data frame was created in

which the images was aligned perfectly as per the labels. This convoluted neural network was created with the three input channels and the six output channels. Further to improve the model performance three CNN techniques was applied which are vanilla CNN, alexnet CNN and the vgg16 CNN which further improved the performance of the model even more as compare to the rest four models.

Results

Models	Accuracy before grid search	Accuracy after grid search
SVM-SIFT	42%	41%
SVM-HOG	64%	63%
MLP-SIFT	42%	41%
MLP-HOG	51%	58%

Table. 1 Performance of the models before and after grid search

As can be seen in the table1. Prior to the implementation of the grid search, the SVM-SIFT model had an accuracy of only 41%. When hyperparameters like gamma and C were introduced, the accuracy was only slightly impacted in a negative way. In a similar manner, the application of the hyperparameters resulted in a drop in the accuracy of the SVM-HOG. After applying the hyperparameters, the accuracy of the model was shown to decline in the MLP-SIFT model, following the same pattern as observed in the SIFT model. Before performing the grid search, the accuracy of the MLP-HOG model was only 51%; after applying the hyperparameters, however, the accuracy increased to 58%. The CNN model performed far better than the other four models combined. The accuracy of the non-specialized CNN model was 70.56 percent when it was trained. After the alex net and vgg16 CNN model were deployed, the accuracy started growing, and for vgg16, the accuracy rose to 80.99 percent.

Fig. 1 Facial emotion recognition

As may be seen in Figure 1, the first feeling that was recognised was that of rage. Sadness was identified as the second feeling, followed by a recognition of neutrality as the third feeling. The model has also inspired a great deal of different kinds of emotional responses. The CNN model was developed in order to generate these types of face expressions. In the beginning, the data that would later be used to develop the model were processed. To begin this processing, a data frame was initially generated so that the labels and images could be aligned. Following that, the CNN model was constructed using three input channels, six output channels, and a central node. The data in the CNN model has previously been transformed before it was normalised. The model was educated using all 30 epochs of data. After that, the vgg16 model was implemented in order to further boost the accuracy. The next step is to implement the emotion recognition function that was established in the test file using the video dataset. In that function, the video function Object() { [native code] } was presented to the user for the first time. The data from the video was then transformed into tensors before being normalised after the transformation. After that, a cascade classifier was applied, and this data on the face was cascaded. In the end, the photos are converted into the array, and after assigning colours to the edges, the function was able to recognise the expression. Following that, this function was applied to the vgg16 CNN model while the video was being processed at the same time.

Fig. 2 Facial emotion recognition using SVM-HOG

As can be shown in Figure 2, SVM-HOG is capable of recognising facial expressions of emotion. In order to acquire these results, the function was originally developed for four different models, excluding the CNN model. In the beginning, this function started by collecting the data from the respective directories. After that, it moved on to collecting the labels from the respective directories. Since this function was initially developed for the SIFT models, the SIFT detector was also developed at the same time. The data were then further transformed to the arrays in order to make it easier to work with. This was done after the main points were identified and descriptors were extracted using SIFT. Additionally, the data was partitioned into batches for the purpose of more efficient calculation. The test data, including the orientations and cell counts for each block, were then presented at the HOG side. In order to acquire the findings, the optimal function for was applied to the model that had done the best; in this instance, that model was SVM-HOG. This model was compared to the other three models, with the exception of the CNN model.

Discussion

Within the scope of this investigation, five models were applied to the data set that was retrieved from the RAF database. Recognizing the various expressions that can be conveyed by one's face was the primary purpose of this study. In the beginning, the data set was split into train data and test data, with 80 percent of the data set being assigned to the train data and the other 20 percent being assigned to the test data. The first model that was used was SVM-SIFT, but it had a poor performance compared to other models, so it won't be used again in future work. This model demonstrated time-consuming abilities and also provided less accuracy compared to other models, which could result in the loss of several different aspects. In addition to this, the performance of the model deteriorated once the grid search was applied.

The results of the second model, which was an SVM-HOG combination and which performed exceptionally well, were taken into consideration. However, the accuracy of this model reduced slightly once the grid search was applied to it. This was the case for all models. However, the most significant drawback of the SVM-HOG model is that it required a longer amount of time than the other models.

The third model, which was called MLP-SIFT, had performance that was comparable to that of the first model, but it required a significantly less amount of time compared to the first two models. As a result, this model should be taken into consideration for usage in future work because MLP-SIFT has the ability to do high-level jobs in a very short amount of time, despite the fact that it offers less accuracy.

When compared to the other models, the MLP-HOG, which was the fourth model, had normal performance. The most significant benefit of this model is that, after applying the grid search, its accuracy improved, and, in addition, the amount of time required by it for training was significantly less than that required by the first model. Nevertheless, the accuracy of this model can be improved even further by employing additional hyperparameters.

Convolutional neural networks, the fifth model, fared exceptionally well in comparison to the previous four models. In terms of the CNN's fundamental model, it achieved an accuracy of 70.56 percent and required less than 60 seconds for training. In addition, when several distinct types of CNNs were used, the accuracy increased to levels higher than 80 percent. However, the amount of time required for vgg16 CNN was significantly longer than that required for regular CNN, which is an important consideration.

References

[1] Q. Li and X. Wang, "Image Classification Based on SIFT and SVM," 2018 IEEE/ACIS 17th International Conference on Computer and Information Science (ICIS), 2018, pp. 762-765, doi: 10.1109/ICIS.2018.8466432.

[2] N. Nguyen, D. Bui and X. Tran, "A Novel Hardware Architecture for Human Detection using HOG-SVM Co-Optimization," 2019 IEEE Asia Pacific Conference on Circuits and Systems (APCCAS), 2019, pp. 33-36, doi: 10.1109/APCCAS47518.2019.8953123.

[3] M. Moradi, M. A. Poormina and F. Razzazi, "FPGA Implementation of Feature Extraction and MLP Neural Network Classifier for Farsi Handwritten Digit Recognition," 2009 Third UKSim European Symposium on Computer Modeling and Simulation, 2009, pp. 231-234, doi: 10.1109/EMS.2009.13.

[4] A. K. Sah, S. Bhowmik, S. Malakar, R. Sarkar, E. Kavallieratou and N. Vasilopoulos, "Text and non-text recognition using modified HOG descriptor," 2017 IEEE Calcutta Conference (CALCON), 2017, pp. 64-68, doi: 10.1109/CALCON.2017.8280697.