End-to-End Multimodal Emotion Recognition by Using Deep Neural Networks *

Zhifeng Wang[u6068466]

Research School of Computer Science, Australian National University, Australia
u6068466@anu.edu.au

Abstract. Emotion recognition plays an important role in HCI system. The process includes finding important parts of facial regions and classifying them into different emotion classes. In this paper, we create a Multimodal CNN based neural network and pruning based neural network for emotion recognition. We demonstrated our result on the benchmark dataset- SFEW dataset. We report an Average Precision for CNN based neural network of 31% across the 7 categories, which has 12% improvement than previous methods.

Keywords: Emotion recognising · Pruning neural network · Multimodal CNN

1 Introduction

Automatically perceiving and recognizing the emotions of human has become the main part of human-computer interaction[11]. Its associated research includes a multidisciplinary enterprise which includes speech analysis, linguistic psychology, learning theory and robotics. A computer with powerful emotion recognition ability will be able to interact and understand human more naturally. Many real world applications such as affect-aware game development and commercial call centre will benefit from such emotion recognition intelligence.

The possible inputs for emotion recognition include different kinds of signals such as text, audio, bio signals and image. For vision based emotion recognition, the visual information can be used such as including face[13], body and pose[14], text[15], speech[16]. However, facial expression is the most important information for analysing human emotion. Despite the continuous research on emotion recognition, an accurate emotion recognition under different environment is still challenge. Many early emotion recognition dataset[1–4] is collected in the "lab controlled" environment where the target person were asked to generate certain facial expressions. These designed facial expression will result in different visual appearance when comparing with the natural environment[6]. Therefore, it is not a good representation of natural facial expression.

Recent advances in the emotion datasets focus on more spontaneous facial expressions. The Static Facial Expressions in the Wild (SFEW) dataset[6] and Acted Facial Expressions in the Wild (AFEW) dataset[17] were collected to mimic more natural environment which include 7 basic emotion categories- Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise[6]. The AFEW is collected from video but AFEW is static dataset from AFEW. Both of the datasets are collected spontaneously and can mimic the facial expression in natural environment. In the Figure 1, it shows the collected sample facial expression images in the SFEW dataset[6]. The images not only include the face and body information and also they include a lot of background information. All the facial expression is in a natural environment. In this paper, we transfer the image information to first 5 principal components of local phase features and principal components of pyramid of histogram of gradients features. There are 675 input information with 7 emotion categories– Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise.

In this paper, our goal is that we focus on using LBP features and PHOG features to recognise emotion on SFEW dataset with pruning neural networks and End-to-End Multimodal CNN based neural network. We compare their performances on the SFEW dataset.

2 Data preprocessing

In the SFEW emotion dataset, there are 675 input features and 1 label column and 5 LBP features and 5 PHOG features. The box plot in the Figure 2 can show the distribution of the original data. The values of original data range from -0.05 to 0.15 and the centers of the data values is around 0, which is not good for neural network to learn. So, I use the Min-Max Normalization to normalize the original data. The Min-max

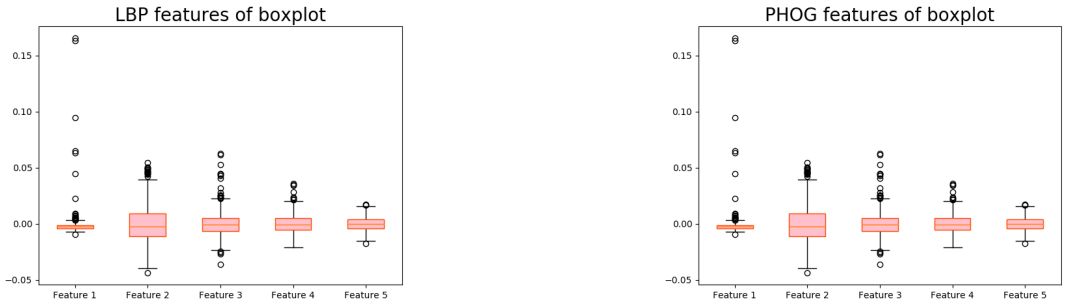
* Supported by Research School of Computer Science, Australian National University

Fig. 1. The collected sample facial expression images in the SFEW dataset[6]. In the SFEW dataset, the facial expression image is spontaneously collected which is used to mimic facial expression in natural environment.

normalization can performs a linear alteration on the original data. The values are normalized within the given range. The benefit of Min-Max normalization is that all the values can be annealed within certain range. The Min-Max Normalization formula[7] can be show in Equation 1:

$$X_{norm} = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{1}$$

After normalizing the original data using Equation 1, the distribution of SFEW data can be shown in Figure 3.



(a) LBP data (b) PHOG data

Fig. 2. The original data distribution of SFEW dataset

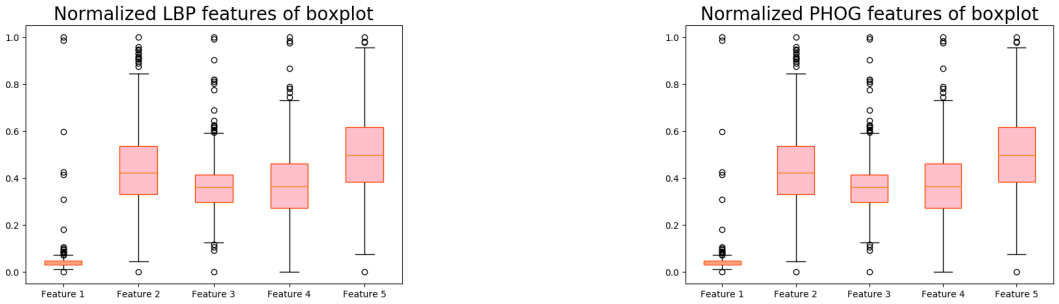
In the Figure 3, we can see that the values of LBP features and PHOG features range from 0 to 1. The center of the LBP features and PHOG features is not near 0, which is good for the neural network to learn. The center for Feature 2, Fearture 3, Fearture 4, Feature 5 is around 0.4.

3 Methods

In this section, the pruning methods for shallow neural networks and the CNN based neural network will be introduced.

3.1 Pruning methods for shallow neural network

This section will focus on pruning trained networks by using distinctiveness of the output from the hidden layer. The distinctiveness of hidden units can be determined from the unit output activation vector. This vector will represent the functionality of the hidden unit. The similarity of pairs of vectors can be calculated by the angle



(a) LBP data (b) PHOG data

Fig. 3. The original data distribution of SFEW dataset

between them in each pattern. Using Cosine similarity to calculate the score of similarity[8]. The similarity formula can be shown in Equation 2, by using this formula, we can calculate the vector angle.

$$similarity(A, B) = \frac{A * B}{||A|| \times ||B||} \tag{2}$$

, where A and B is the vector of outputs of hidden units .

The neural network architecture can be shown in Figure 4, There are three layers in total. The first layer is input layer, the second layer is hidden layer, the third layer is output layer. The dimension of input layer is 5 which is corresponding to LBP and PHOG features dimensions. The output layer dimension is 7 which is corresponding to 7 emotion categories. The number of hidden layers is range from 6 to 16. We analyse the impacts of different number of hidden units on the accuracy.

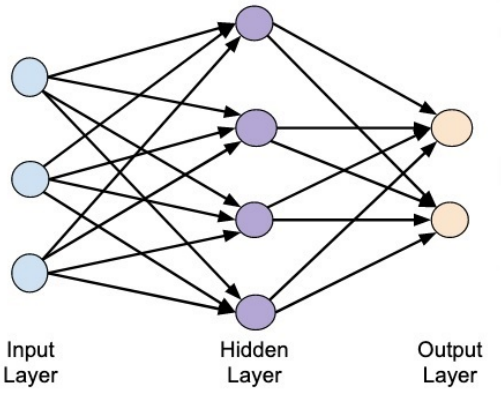


Fig. 4. The neural network architecture include three layers- input layer, hidden layer and output layer

In the Table 1, there are 10 patterns with six hidden units output. It is possible to discover the pairs of hidden units with similar functionality because the number of pattern and hidden units are very small. But for large number of patterns and hidden units, it will become difficult to find the the pairs of hidden units with similar functionality. So we need to calculate the vector angles for the six hidden units, by using the vector angles, we can remove the similar hidden units[9]. The vector angles for these six hidden units are shown in the Table 2.

In the Table 2, we can know that the smaller the angle vector is, the more similar the vector is. The pair of 3 and 4 hidden units have the largest similarity because these pair has the smallest vector angle which is 8.1^o. So we need to remove one of the pair hidden units and ensure that the remaining hidden units have different functionality.

Our neural network pruning algorithm can be described in Algorithm 1: The first step is to train the network one epoch. The second step is to calculate the similarity for each pair of hidden unit and remove similar hidden

unit by a certain angle. So the hidden unit will have different functionalities. The third step is to create a new network without similar hidden unit. Then training the neural network for 200 epochs until it converged.

Table 1. Six hidden unit activations by pattern.

Pattern	1	2	3	4	5	6
p.000	0.5914	0.3303	0.6544	0.5678	0.4130	0.5453
p.001	0.2656	0.4397	0.6113	0.7139	0.3363	0.3672
p.002	0.4798	0.5740	0.5485	0.5522	0.4949	0.4054
p.003	0.6408	0.3224	0.6555	0.5744	0.4975	0.6006
p.004	0.6717	0.3109	0.6153	0.5543	0.4379	0.5231
p.005	0.2993	0.3330	0.7186	0.7569	0.4357	0.5860
p.006	0.4052	0.5558	0.4583	0.6235	0.4310	0.2651
p.007	0.3316	0.5303	0.5024	0.6139	0.3146	0.2507
p.008	0.2550	0.5957	0.5150	0.6451	0.3460	0.2577
p.009	0.3553	0.4690	0.6804	0.6866	0.4933	0.5542

Table 2. Vector angles for pairs of the hidden units.

Pair of units	Vector angle
1 2	29.2
1 3	18.9
1 4	24.2
1 5	16.0
1 6	17.3
2 3	20.3
2 4	14.9
2 5	18.1
2 6	29.7
3 4	8.1
3 5	8.8
3 6	10.5
4 5	11.8
4 6	18.2
5 6	13.0

Algorithm 1 Neural network pruning algorithm

1: Start train the network for 1 epoch
2: After 1 epoch, calculate the similarity for each pair of hidden unit and remove similar hidden unit by a certain angle
3: Create a new network without similar hidden unit
4: Repeat Step 1 to Step 3 and train the network for 200 epochs until it converged

3.2 Convolution neural network model for LBP and PHOG features

Convolutional neural network (CNN) is the most popular way of extracting features from input data. CNN is a type of Neural Network where the mathematical operation can be used to find the relationship between the data [18] [19]. In this paper, we use 5 layers convolutional neural networks to extract LBP features and PHOG features. The pipeline of the model can be shown in Figure 5 and it is divided into three modules: LBP feature extraction module, PHOG feature extraction module and fusion module. The first module takes the LBP features. The second module takes PHOG features. Finally, the third modules combines these features to

do a fine-grained regression of the two types of emotion representations. Both of the feature extraction modules are based on one dimension CNN [20]. These CNN networks provide a competitive performance although the number of parameters is low. Each network consists of 5 layers, each layer has one convolution with 3 dimension kernals and Maxpooling. The green box in the Figure 5 represent the convolution operation and the red box in the Figure 5 represent the maxpooling operation.

The fusion module consists of two fully connected (FC) layers. The first FC layer is used to reduce the dimensionality of features to 14. The dimension of output for first FC layer is twice the size of the discrete emotion categories. The second fully connected layer is to predict the emotion category and the dimension is 7 which are the same as the number of emotion categories.

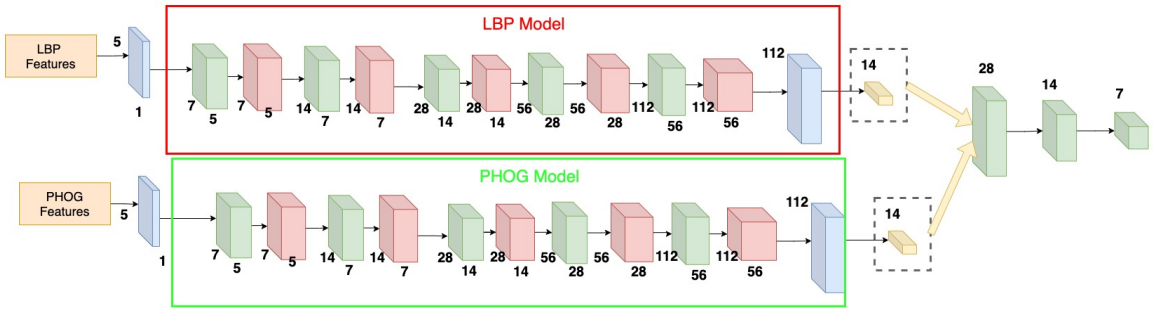


Fig. 5. Proposed end-to-end model for emotion recognition by using LBP features and PHOG features. The model consists of two feature extraction modules and a fusion network for jointly estimating the discrete categories

3.3 Loss function

In the Section 2 data preprocessing part, we found that the data have a lot of noise. So we use Smooth L1 Loss as our model's loss function. The Smooth L1 Loss function is robust to the noise data [5] . The Smooth L1 Loss can be described as following formula:

$$smooth_{L1}(x) = \begin{cases} 0.5x^2 & if|x| \leq 1. \\ y = |x| - 0.5 \; otherwise. \end{cases} \tag{3}$$

4 Experiments and Results

4.1 Training details

In this section, we show the results and perform comparative trial with the baseline. In our experiment, we adopt the following parameters. The input size is 5 which is corresponding to the LBP features dimension and PHOG features dimension. The default hidden size for pruning model= 16 from technique paper[9]. The number of classes is 7 which is corresponding to 7 emotion categories-angry, disgust, fear, happy, neutral, sad, and surprise. For pruning network, because we need to prune the network, so it will need to more time to converge, so the default epochs for pruning network is 200, batch size is 10, learning rate is 0.0001. The CNN based network use the following parameters: the number of epochs is 30, batch size is 10, learning rate is 0.0001. Using Pytorch tool[10] to implement the pruning neural network and CNN based Multimodal neural network.

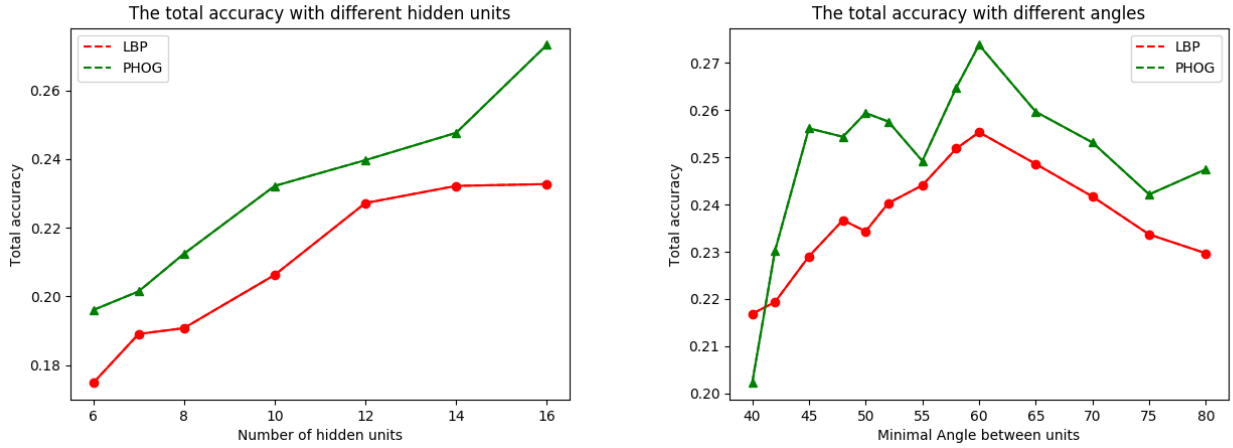
4.2 Evaluation Metrics and methods

We use the standard Average Precision (AP) score to evaluate our results and methods on the SFEW dataset. We compare our results with the baseline model and report our results. The baseline model use SPI method to get the baseline result[9].

4.3 Analysis and Discussion

Comparing with baseline method In the Figure 6, we analyse the relationship between accuracy and the number of hidden units and minimal angle between hidden units. we found that in the Figure 6 (a), we can see that the accuracy increase when the number of hidden units increase and the accuracy of PHOG dataset

is higher than the accuracy of LBP dataset, which means that the more number of hidden units will help the neural network get higher performance on recognising emotions. In the Figure 6 (b), the accuracy increase as the minimal angle between units increase. When the minimal angle between units is 60 degree, the accuracy has the largest number. So, we choose the 60 degree to prune our neural network which has the highest average accuracy. In the PHOG dataset, our network achieves better performance than LBP dataset.



(a) Accuracy and hidden units

(b) Accuracy and minimal angle

Fig. 6. The impacts of hidden units and minimal angle between units

In the Table 3, it gives the accuracy comparing with the SPI protocol baseline accuracy. Comparing with the baseline method, our method for shallow network can have 8% improvement on the average accuracy from 19% to 27%. Our methods by using CNN netowrk can archive 12% improvement on the average accuracy from 19% to 31%. We found that the pruning method can improve the "Angry" emotion from 17% to 38%, the "Happy" emotion from 28% to 36%. At the same time, we compare our methods using different datasets- LBP dataset and PHOG dataset. By using the pruning method, the emotion categories "Disgust", "Fear" and "Surprise" have the highest accuracy on the LBP dataset. The emotion categories "Angry", "Happy" and "Surprise" have the highest accuracy on the PHOG dataset. Comparing with the pruning method, our method by using CNN fusion network can have the highest accuracy. By analysing the accuracy of CNN fusion model, the fusion model can use the both LBP features and PHOG features' advantages. For "Happy" category, the PHOG for CNN model has the highest accuracy, although the accuracy for LBP CNN model for "Happy" category is only 0.27, the CNN fusion model for "Happy" category can reach 45%.

Table 3. Comparing with the SPI protocol baseline accuracy

Emotion	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise	Average Accuracy
SPI Protocol (Baseline)[6]	0.17	0.15	0.20	0.28	0.22	0.16	0.15	0.19
LBP for shallow network (Ours)	0.14	0.34	0.23	0.34	0.20	0.15	0.20	0.23
PHOG for shallow network (Ours)	**0.38**	0.20	0.18	0.36	**0.33**	0.28	0.16	0.27
LBP for CNN model (Ours)	0.14	0.28	0.24	0.27	0.18	0.19	0.11	0.20
PHOG for CNN model (Ours)	0.35	0.23	0.21	**0.46**	0.21	0.15	**0.35**	0.28
CNN fusion model (Ours)	0.29	**0.38**	**0.31**	0.45	0.19	**0.20**	0.33	**0.31**

Analysis and Discussion for fusion CNN model In the Figure 7, we compared the performance for different emotion recognition models in different emotion categories. For single LBP feature model, it performance best in "Fear" emotion category, but in other emotion category, it performance worse than other two models. The fusion model performance best on the "Angry", "Disgust", "Happy" and "Surprise" and fusion model has the highest average accuracy. The fusion model has the highest average accuracy because the fusion part of the

fusion network can merge the LBP features and PHOG features extracted from previous neural network.

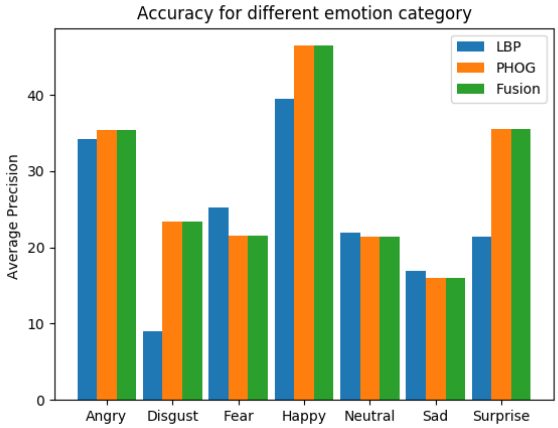
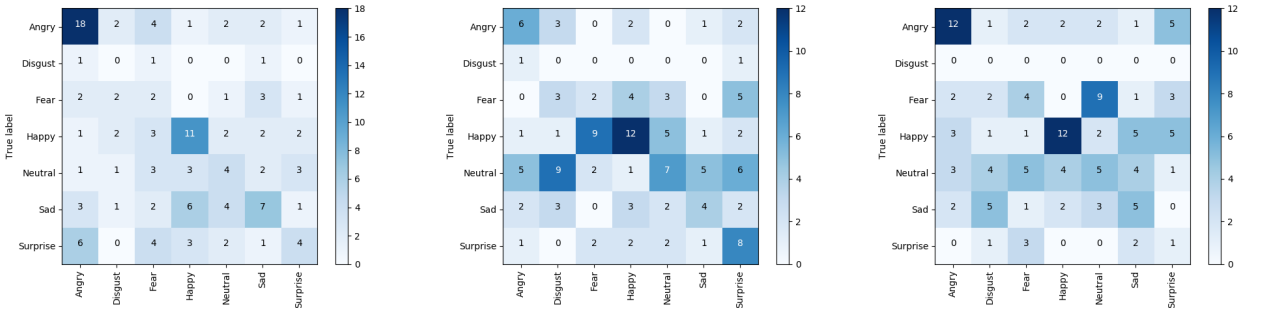


Fig. 7. Accuracy for CNN based model. Proposed end-to-end CNN fusion model for emotion recognition by using LBP features and PHOG features. The model consists of two feature extraction modules and a fusion network for jointly estimating the discrete categories

In the Figure 8, it gives the confusion matrix for single LBP model, single PHOG model and fusion model. In the Figure 8 (a), we found that the single LBP model can distinguish "Angry" with other emotion categories. In the Figure 8 (b), we found that the single PHOG model can distinguish "Happy" with other emotion categories. In the Figure 8 (c), we found that the fusion model can distinguish "Angry" and "Happy" with other emotion categories. But our CNN fusion model has limited ability to distinguish "Disgust" with other emotion categories.

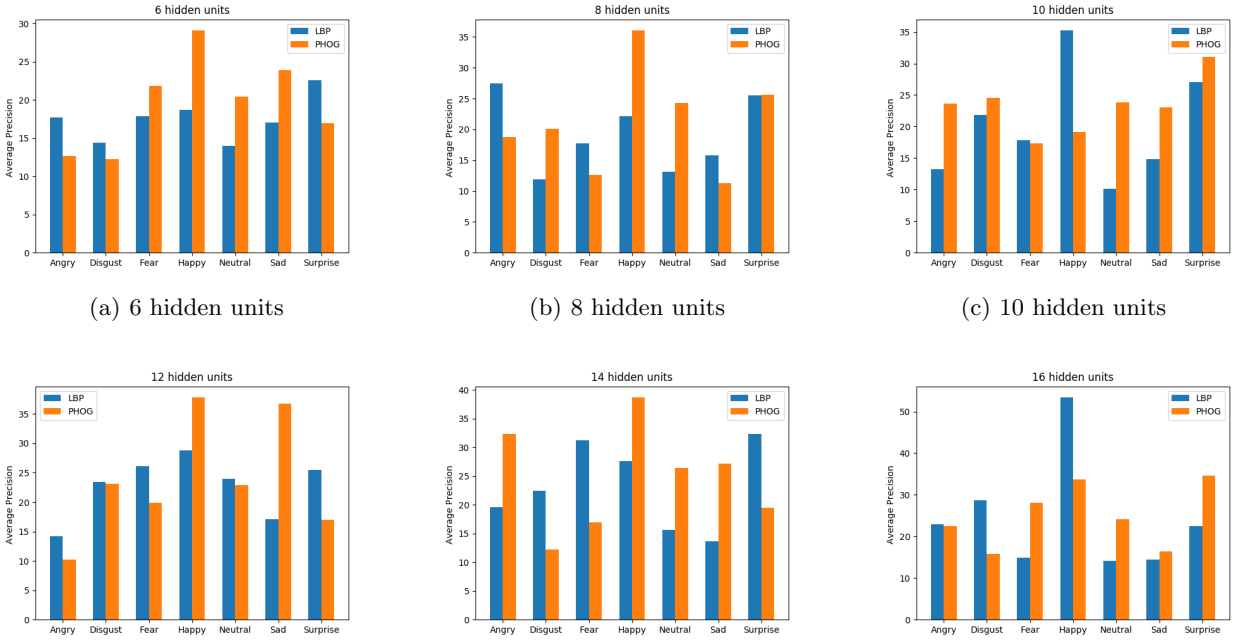


(a) Confusion matrix for LBP model (b) Confusion matrix for PHOG model (c) Confusion matrix for fusion model

Fig. 8. Confusion matrix for CNN based model. Confusion matrix for singleLBP model, single PHOG model and fusion model

Limitations for pruning network and CNN models In the Figure 9, we show the Average Precision for shallow pruning network on LBP data and PHOG data for 7 emotion categories- Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise with different hidden units, which is corresponding to different compression ratio. In the Figure 6 (a), the average accuracy for all the emotions increase when the number of hidden units increase. In the Figure 7, we found that our fusion CNN performance better than the shallow pruning network which has the highest average accuracy and has %12 improvement than the baseline method. Comparing with the baseline method, our shallow pruning network and fusion CNN method has achieved greatly improvement.

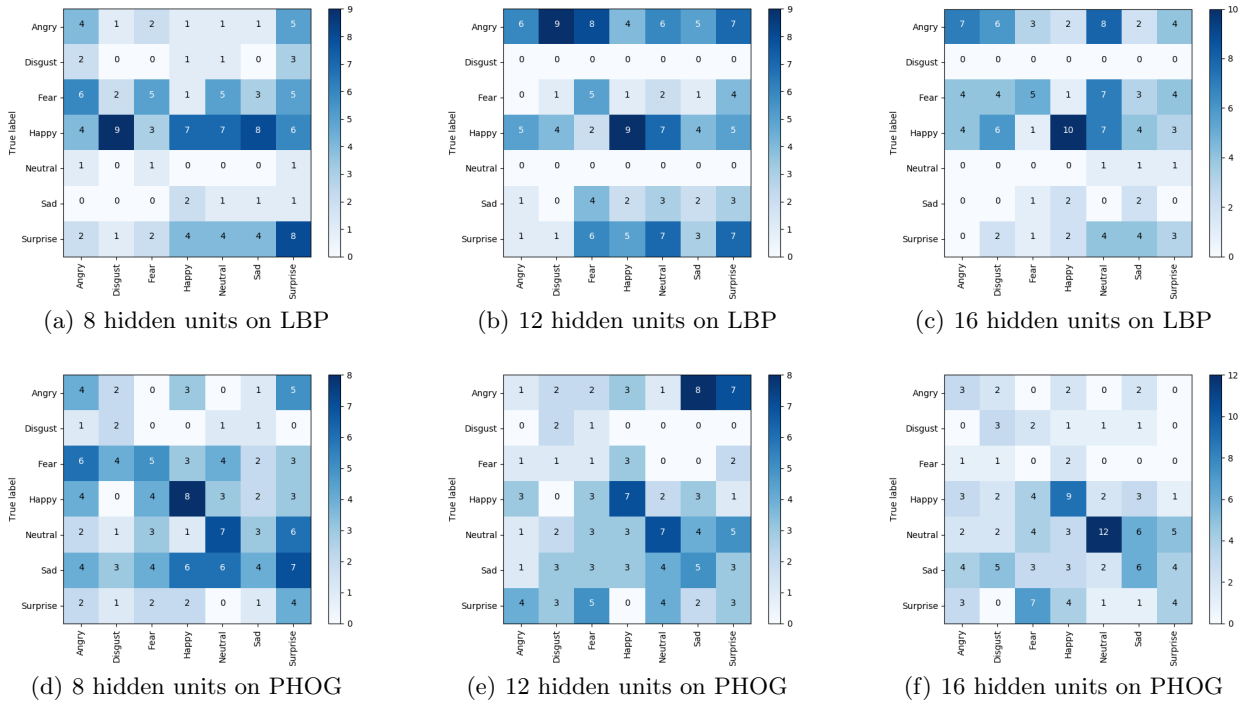
However, both our shallow pruning network and fusion CNN network has some limitations. In the Figure 9, for "Happy" emotion on PHOG dataset, the accuracy of "Happy" category first increase then decrease when the number of hidden units increase. "Happy" category on PHOG dataset has the lowest number in 16 hidden units, which means our model has limitation for distiguishing "Happy" category with other emotion categories,



(a) 6 hidden units (b) 8 hidden units (c) 10 hidden units

(d) 12 hidden units (e) 14 hidden units (f) 16 hidden units

Fig. 9. Accuracy for pruning method model. The Average Precision on LBP data and PHOG data for 7 emotion categories- Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise



(a) 8 hidden units on LBP (b) 12 hidden units on LBP (c) 16 hidden units on LBP

(d) 8 hidden units on PHOG (e) 12 hidden units on PHOG (f) 16 hidden units on PHOG

Fig. 10. Confusion matrix for pruning method model. The confusion matrix on LBP data and PHOG data for 7 emotion categories- Angry, Disgust, Fear, Happy, Neutral, Sad and Surprise

and our model is not robust for certain noise, although the average accuracy on PHOG is the highest. In the Figure 10 (f), the confusion matrix shows that our pruning model have limited ability to distinguish the "Anger" and "Surprise", "Surprise" and "Sad". In the Figure 8 (c), we found that the CNN based fusion model can distinguish "Angry" and "Happy" with other emotion categories. But our CNN fusion model has limited ability to distinguish "Disgust" with other emotion categories. This low accuracy is attribute to the complex nature of condition in the SFEW database[6]. Our shallow neural network and CNN fusion models are not robust for uncontrolled environment experiments. The other reason is that our model use LBP features and PHOG

features which have been preprocessed and some local and global information in the images has lost. The LBP features and PHOG features is hard for our model to be used to infer the person's emotion.

5 Conclusion and Future Work

We have demonstrated that using pruning network and fusion CNN network can improve the ability of recognising emotion states. Our proposed methods achieved an excellent result on SFEW dataset. However, the SFEW dataset is uncontrolled environment dataset, the confusion matrix shows that our shallow model has limited ability for distinguishing certain emotions. Our future work is to use image-based dataset to improve the ability for distinguishing certain emotion states. In this paper, we used the LBP and PHOG features dataset and analyse the performance on these two datasets. However, our pruning network and fusion CNN network has limited ability to extract useful information by just using LBP features and PHOG features. In the future work, we need to combine LBP and PHOG features dataset and image datasets and take advantage of image information and PCA features information to training our model, and improve the ability for recognising emotion states.

References

1. P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews. The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on, pages 94–101. IEEE, 2010.
2. V. Ojansivu and J. Heikkila. Blur insensitive texture classification using local phase quantization. In Image and signal processing, pages 236–243. Springer, 2008.
3. T. B anziger and K. R. Scherer. Introducing the geneva multimodal emotion portrayal (gemep) corpus. Blueprint for affective computing: A sourcebook, pages 271–294, 2010.
4. M. S. Bartlett, G. C. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan. Automatic recognition of facial actions in spontaneous expressions. Journal of multimedia, 1(6):22–35, 2006.
5. Near-duplicated Loss for Accurate Object Localization
6. Dhall, Abhinav, et al. "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark." 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops). IEEE, 2011.
7. Saranya, C., and G. Manikandan. "A study on normalization techniques for privacy preserving data mining." International Journal of Engineering and Technology (IJET) 5.3 (2013): 2701-2704.
8. Lahitani, Alfirna Rizqi, Adhistya Erna Permanasari, and Noor Akhmad Setiawan. "Cosine similarity to determine similarity measure: Study case in online essay assessment." 2016 4th International Conference on Cyber and IT Service Management. IEEE, 2016.
9. Gedeon, T. D., and D. Harris. "Progressive image compression." [Proceedings 1992] IJCNN International Joint Conference on Neural Networks. Vol. 4. IEEE, 1992.
10. Ketkar, Nikhil. "Introduction to pytorch." Deep learning with python. Apress, Berkeley, CA, 2017. 195-208.
11. Fragopanagos, Nickolaos, and John G. Taylor. "Emotion recognition in human–computer interaction." Neural Networks 18.4 (2005): 389-405.
12. Karras, Tero, et al. "Audio-driven facial animation by joint end-to-end learning of pose and emotion." ACM Transactions on Graphics (TOG) 36.4 (2017): 1-12.
13. Lasri, A. R. Solh and M. E. Belkacemi, "Facial Emotion Recognition of Students using Convolutional Neural Network," 2019 Third International Conference on Intelligent Computing in Data Sciences (ICDS), Marrakech, Morocco, 2019, pp. 1-6, doi: 10.1109/ICDS47004.2019.8942386.
14. F. Ahmed, A. S. M. H. Bari and M. L. Gavrilova, "Emotion Recognition From Body Movement," in IEEE Access, vol. 8, pp. 11761-11781, 2020, doi: 10.1109/ACCESS.2019.2963113.
15. Leung, J. K., Griva, I., and Kennedy, W. G. (2020). Text-based Emotion Aware Recommender. arXiv preprint arXiv:2007.01455.
16. Issa, Dias, M. Fatih Demirci, and Adnan Yazici. "Speech emotion recognition with deep convolutional neural networks." Biomedical Signal Processing and Control 59 (2020): 101894.
17. Kossaifi, Jean, et al. "AFEW-VA database for valence and arousal estimation in-the-wild." Image and Vision Computing 65 (2017): 23-36.
18. C. J. L. Flores, A. E. G. Cutipa and R. L. Enciso, "Application of convolutional neural networks for static hand gestures recognition under different invariant features," 2017 IEEE XXIV International Conference on Electronics, Electrical Engineering and Computing (INTERCON), Cusco, 2017, pp. 1-4.
19. J. Shijie, W. Ping, J. Peiyi and H. Siping, "Research on data augmentation for image classification based on convolution neural networks," 2017 Chinese Automation Congress (CAC), Jinan, 2017, pp. 4165-4170.
20. J. Alvarez and L. Petersson, "Decomposeme: Simplifying convnets for end-to-end learning," CoRR, vol. abs/1606.05426, 2016.