Emotion Detection using Periocular Region: A Cross-Dataset Study

Narsi Reddy
Dept. of Computer Science and Electrical Engineering
University of Missouri at Kansas City, MO, USA-64110
sdhy7@mail.umkc.edu

Reza Derakhshani
Dept. of Computer Science and Electrical Engineering
University of Missouri at Kansas City, MO, USA-64110
derakhshanir@umkc.edu

***Abstract*—Many computer vision methods have been proposed for the affective assessment using facial expressions from full-face images. In many use cases, however, only the ocular region may be available due to the application of masks, clothing items, or privacy issues. In this paper, we show the utility of a robust and yet light deep learning model for ocular affect assessment using cross-dataset evaluation, where we train the model on one dataset and perform testing on another dataset with different demographics. We compare a MobileNet-V2 deep learning model, using transfer learning, with a more traditional method using histogram of oriented gradients (HOG) features with support vector machine (SVM) classifier. Experiments were conducted on the FACES dataset for training with six facial expressions and tested on the more diverse Chicago faces dataset(CFD) to show how evaluated models generalize not only in cross-dataset evaluation but also in the presence of new ethnicities not present during training. The experimental results show that the deep learning model can provide an average accuracy of 76.77% overall facial expressions when compared to HOG and SVM's 62.47% for this challenging cross-dataset emotion assessment using only eye regions.**

***Index Terms*—Expression Recognition, Emotion Recognition, Deep Learning, Eye Region, Periocular**

I. Introduction

Automatic assessment of emotion from facial expressions has numerous applications, including medical psychology, special effects and animations, security, human-computer interface (HCI), and marketing. Research interest in studying automatic emotion recognition from facial expressions dates back to early 1970*s* [1], [2]. In [2], Ekman and Friesen defined six basic emotions based on facial expressions, which are: anger, disgust, fear, happiness, sadness, and surprise.

A detailed survey of facial expression recognition systems are provided in [3]–[5]. Most earlier techniques proposed handcrafted feature extraction techniques such as local binary patterns(LBP) [6], local phase quantization (LPQ) [7], histogram of oriented gradients (HOG) [8], Gabor filters [9], and scale-invariant feature transform (SIFT) [10]. For recognizing emotion, feature extraction models are coupled with classification techniques such as support vector machines (SVM) and decision trees with boosting. With advancements in deep learning technology, many recent papers have turned to convolutional neural networks(CNN) [11]–[13] based methods achieving state of the art performance.

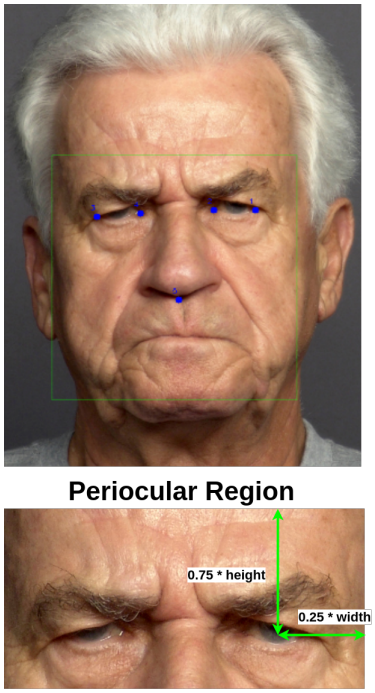

Fig. 1. Extracting periocular region from the full facial images.

Most of the proposed methods are designed to recognize emotion from full-face images. Thus, they may not be able to operate on occluded faces [14], mainly due to surgical masks worn to avoid communicable diseases and air pollution populated cities, or other cultural and religious face coverings. Furthermore, regardless of how a person is holding a mobile device, in most mobile application scenarios, one can assume that at least the eye-band region (also called periocular) is mostly visible to the front-facing camera and thus available for emotion recognition [15].

In [14], Zhang et al., authors evaluate many techniques proposed for full-face image emotion assessment on partial face images. They show that the performance of the eye region in emotion detection is higher than in other areas of the face. Alonso-Fernandez et al., in [16], performed a feasibility study on the periocular region for expression recognition using a fusion of several handcrafted feature descriptors such as

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

LBP, HOG, and Gabor with linear support vector machine (SVM) classifiers. Using the Extended Cohn-Kanade Dataset (CK+) [17], the authors show an overall accuracy of 78% on eight facial expressions. In [18], Elmar Langholz proposed to infer the overall affective state of a person from just eye region using CNN based methods. Most of these reports on the periocular expression detection are developed and tested on the same dataset, forgoing non-heterogeneous experiments on how the periocular region-based expression detection could work across-datasets where models trained on one dataset and tested on another.

In this paper, we compare the performance of handcrafted features with a deep learning based model framework for emotion recognition using the periocular region of the face (the band encompassing the area from lower eyelids to the upper eyebrow region), with high generalization power as shown in our cross-dataset evaluation. To the best of knowledge, this is the first study to perform cross-dataset performance evaluation on expression recognition from the periocular region. More specifically, we present:

1) For handcrafted features, we created a well-known pipeline using HOG feature descriptor along with SVM classifiers
2) For deep learning based model, a MobileNet-V2 [19] model, which is an efficient architecture designed to work efficiently with high inference speed even on mobile devices.
3) Cross-dataset and open set evaluation, by training model on the FACES [20] dataset with 6 expression classes, and testing on complete Chicago faces dataset [21] with non-overlapping and ethnically diverse participants.
4) Since the training dataset is relatively small to train the deep learning model from scratch, we show a successful application of transfer learning on the MobileNet-V2 model, which was pre-trained on the ImageNet [22] dataset and further tuned for expression recognition using the periocular region.

The rest of this paper is as follows: Section II details the learning based models evaluated. In Section III, datasets, experimental protocol, and experimental results are presented. Finally, conclusions are drawn in Section IV.

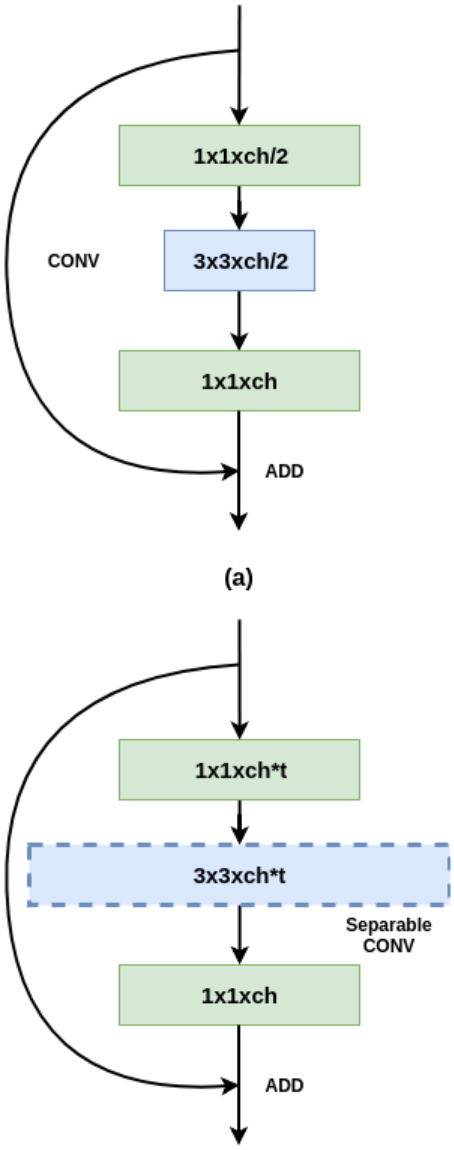

Fig. 2. (a) Original ResNet architecture with 3×3 convolutional operations on squeeze features channels. (b) MobileNet-V2 architecture with inverted residuals where separable 3×3 convolutions are applied to expanded feature channels.

II. Evaluated Models

A. Handcrafted Features based method ($HOG + SVM$):

For handcrafted features based method, we computed dense HOG feature descriptors with SVM classification. For both protocols from Section III-B, we incorporated a grid search to find the best combination of global parameters for both HOG descriptors and SVM classifiers ($HOG + SVM$).

To find the best parameters for dense HOG descriptors (i.e., evaluated over a regular image grid) with 8 orientations, we tested three cell sizes from 8×8 pixels to 32×32 pixels. Finally, features were normalized using $l2$ normalization. In the case of the SVM classifier, we used a one-vs-all decision function. We searched for the best combination with four kernels: linear, radial basis functions (RBF), a polynomial of order 2, and polynomial of order 3, and five C constant values from 0.0001 to 100.0.

As a result of this exhaustive search on data splits from Section III-B for the best combination of meta parameters, we found HOG descriptors with 8 orientations, 8×8 cell size with $l2$ normalization to be the best. For classification, one-vs-all SVM classifier with polynomial of order 3 kernel with $C = 0.1$ constant emerged as the best choice.

B. Deep learning based model (MobileNet-V2):

For expression recognition from the periocular region, we propose MobileNet-V2 architecture [19] for its lower computational cost, making it suitable for deployment on energy

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE I
CNN MODEL BASED ON MOBILENET-V2 ARCHITECTURE. *Note: The input shapes are described in height × width × channels.*

Input	Layer
224×224×3	ConvBNReLU (3×3×32, stride-2)
112×112×32	1 × InvertedResidual(t = 1, ch = 16)
112×112×16	2 × InvertedResidual(t = 6, ch = 24, stride = 2)
56×56×24	3 × InvertedResidual(t = 6, ch = 32, stride = 2)
28×28×32	4 × InvertedResidual(t = 6, ch = 64, stride = 2)
14×14×64	3 × InvertedResidual(t = 6, ch = 96, stride = 1)
14×14×96	3 × InvertedResidual(t = 6, ch = 160, stride = 2)
7×7×160	1 × InvertedResidual(t = 6, ch = 320, stride = 1)
7×7×320	ConvBNReLU (3×3×1280)
7×7×1280	Global Average Pooling
1×1×1280	Dropout(50%)
1×1×1280	Conv 1×1×K (output)

and memory-constrained computing environments such as those found in mobile devices. MobileNet-V2 architecture is based on inverted residuals, where unlike residual layers from ResNet architecture, the separable 3×3 convolutional features are applied on expanded feature channels, as shown in Figure 2. Table I defines MobileNet-V2 architecture, where K is the number of target classes. During training, a dropout regularization of 50% is applied to the layer preceding the classification layer to avoid over-fitting and to improve the generalization power of the model.

In **transfer learning**, a model developed for a specific task is fine-tuned to adapt to a new task domain, usually using far fewer samples than what is needed for training a deep learning architecture from scratch. This is achieved by mostly training the task-specific latter layers in the network since earlier layers usually learn standard lower-level features and concepts common to many tasks. Accordingly, for transfer learning to do expression recognition from the periocular region, we used the MobileNet-V2 model pre-trained on the very large ImageNet dataset for large scale visual recognition. The final classification layer with $K = 1000$ classes from ImageNet is replaced with a new layer of $K = 6$ target emotion classes as the output for our experiments.

For fine-tuning the model, we used Adam [23] optimizer with a batch size of 32 images and a learning rate of $lr = 10^{-4}$ for all the layers with pre-trained weights from ImageNet. For the final classification layer, the learning rate was increased to $lr = 10^{-2}$ for better parameterization and adaptation to the new domain. The remaining parameters of Adam optimizer were left at their default values.

III. EXPERIMENTAL RESULTS

In this section, first, we provide a brief overview of the databases used in this work, followed by a brief explanation of the periocular region of interest extraction and reprocessing.

TABLE II
NUMBER OF SAMPLES AVAILABLE IN FACES AND CFD DATASET FOR FACIAL EXPRESSIONS: NEUTRAL, SADNESS, DISGUST, FEAR, ANGER, AND HAPPINESS.

Expression	FACES	CFD
Neutral	342	597
Sadness	342	-
Disgust	342	-
Fear	342	149
Anger	342	154
Happiness	342	307
Total	**2,052**	**1,207**

Next, we detail the training and testing protocols for our evaluated methods.

A. Datasets:

FACES [20]: This dataset contains high-quality face images collected from 171 Caucasian volunteers. The dataset is divided into three age groups, young ($n = 58$), middle-aged ($n = 56$), and older ($n = 57$), all expressing six facial expressions: neutral, sadness, disgust, fear, anger, and happiness. The dataset comprises two samples per expression per person with a total of $2,052$ images.

CFD [20]: The Chicago Face Database consists of face images from 597 volunteers with different ethnicity, displaying one of five different facial expressions for fear, anger, happy close-mouthed, happy open-mouthed, and neutral. As our work is focused on expression recognition using only the periocular region, we combined both *Happy close-mouthed* and *Happy open-mouthed* classes into one class *Happy* to overlap with FACES dataset labels. Thus in total, we ended up with four different facial expressions classes, with a total of $1,207$ facial images, for the CFD dataset.

Table II shows the distribution of the sample in both FACES and CFD datasets for available facial expressions. Given its more even distribution of samples and additional facial expression classes, we used the FACES dataset for both training and testing. As for the CFD dataset, we used it only for testing. Figure 3, shows sample full-face images and their corresponding periocular images from both FACES and CFD datasets.

To generate the periocular region from face images, we used the Dlib library [24] for face detection and facial landmark localization. For eye localization, Dlib's 5-point face landmarks are used. Periocular region crops were generated such that both eye centers are at 75% of the height of the image from the top. From the sides, both eye centers were at 25% of the crop width, as shown in Figure x. All the crops were resized to 224×224 for our deep learning method. For the HOG descriptor with the SVM classifier, we resized images to 64×128.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

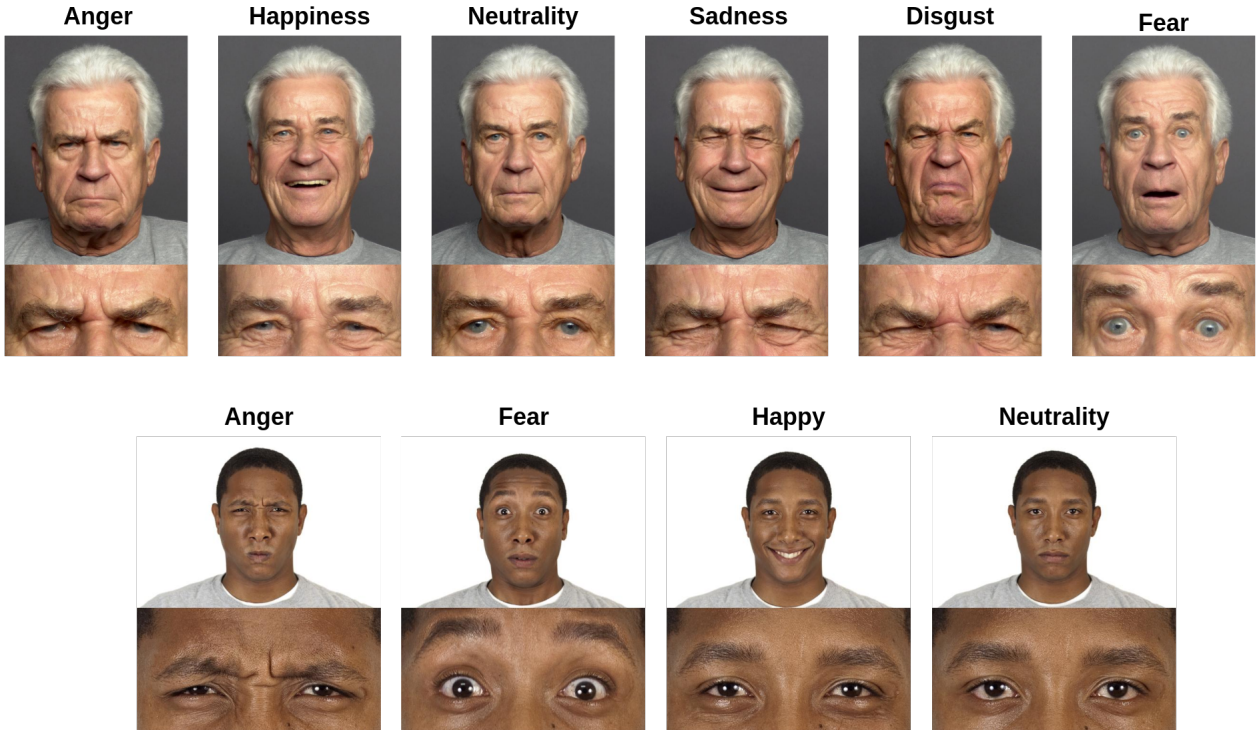

Fig. 3. (First Row) FACES dataset samples - a volunteer is showing six facial expressions: anger, happiness, neutral, sadness, disgust, and fear. (Second Row) Chicago Face Database samples - a volunteer is showing five facial expressions: anger, fear, happiness open-mouthed, and neutral

B. *Protocol:*

We evaluated the performance of both learning models from Section II under two protocols.

First **open-set environment**, where the user identities present in the training set are not present in the testing set. For this setting, we used the FACES dataset where volunteers from each age group are randomly divided into three data splits for training (60%), validation (10%), and testing (30%). The deep learning model was fine-tuned on training splits for 20 epochs with a batch size of 32. Early stopping is considered using classification accuracy of the model on validation data split, and then the best weights were considered for evaluation on testing splits.

For our experiments, we generated the random split 10×. We took the average of testing accuracies to smooth the stochastic training variations, and better show the overall performance of the evaluated models under the open set protocol.

Second **cross-dataset protocol**, where we train the model on one dataset and test the performance on a different dataset, is especially challenging and interesting since it adds heterogeneity of a different testing dataset on top of the open-set protocol described above. Here we used the FACES dataset for training and evaluated the resulting model on the whole CFD dataset. To train the models, the FACES dataset was divided into a training split (90%) and a validation split (10%). Similar to the open-set experiment, for the deep learning model, early stopping on validation data split with the best classification accuracy is considered.

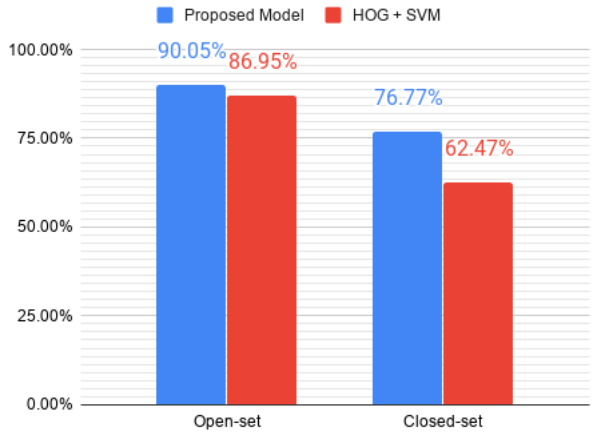

Fig. 4. The average accuracy of all facial expressions for the MobileNet-V2 model and $HOG + SVM$ model in both open-set evaluation and cross-dataset evaluation.

In this cross-dataset experiment, even though the CFD dataset had no data with facial expression *Sadness* and *Disgust*, we trained the model on all six facial expressions from FACES dataset to show how well our evaluated models generalizes on a new dataset especially with a more diverse ethnic representation not available during training.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

TABLE III
CONFUSION MATRIX SHOWING CLASSIFICATION ACCURACY FOR EACH FACIAL EXPRESSION CLASS USING THE MOBILENET-V2 MODEL AND $HOG + SVM$ METHOD IN OPEN-SET EVALUATION ON THE FACES DATASET.

MobileNet-V2	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	**91.26%**	9.62%	0.00%	0.00%	1.92%	7.69%
Disgust	3.88%	**83.65%**	0.00%	7.69%	0.00%	2.88%
Fear	0.00%	0.00%	**100.00%**	0.00%	0.00%	0.00%
Happiness	0.00%	3.85%	0.00%	**90.38%**	1.92%	2.88%
Neutral	0.97%	0.00%	0.00%	0.96%	**90.38%**	1.92%
Sadness	3.88%	2.88%	0.00%	0.96%	5.77%	**84.62%**

HOG + SVM	Anger	Disgust	Fear	Happiness	Neutral	Sadness
Anger	**76.42%**	6.60%	0.00%	0.00%	0.00%	3.77%
Disgust	12.26%	**84.91%**	0.00%	7.55%	0.00%	1.89%
Fear	0.94%	0.00%	**98.11%**	0.94%	2.83%	0.00%
Happiness	0.00%	6.60%	0.00%	**84.91%**	0.94%	0.94%
Neutral	0.94%	0.00%	0.00%	2.83%	**89.62%**	5.66%
Sadness	9.43%	1.89%	1.89%	3.77%	6.60%	**87.74%**

TABLE IV
CONFUSION MATRIX SHOWING CLASSIFICATION ACCURACY FOR EACH FACIAL EXPRESSION FOR THE MOBILENET-V2 AND $HOG + SVM$ METHOD IN CROSS-DATASET EVALUATION WITH MODELS TRAINED ON THE FACES DATASET AND TESTED ON THE CFD DATASET.

MobileNet-V2	Anger	Fear	Happiness	Neutral
Anger	**30.92%**	0.00%	0.00%	0.00%
Disgust	17.76%	1.35%	0.00%	0.00%
Fear	1.97%	**79.05%**	0.98%	0.67%
Happiness	5.92%	5.41%	**63.73%**	4.03%
Neutral	27.63%	6.76%	33.66%	**94.62%**
Sadness	15.79%	7.43%	1.63%	0.67%

HOG+SVM	Anger	Fear	Happiness	Neutral
Anger	**27.63%**	0.00%	0.00%	0.00%
Disgust	48.68%	3.38%	0.98%	1.51%
Fear	2.63%	**82.43%**	2.61%	7.06%
Happiness	17.76%	6.08%	**87.91%**	32.94%
Neutral	0.66%	1.35%	8.50%	**53.95%**
Sadness	3.95%	7.43%	0.33%	4.87%

C. Results

Figure 4 show the average accuracy for all facial expressions for the deep learning model (MobileNet-V2), along with the handcrafted features ($HOG+SVM$) model, under the open-set protocol (on FACES dataset) and cross-dataset protocol (with models trained on FACES dataset and tested on CFD dataset). Overall, the MobileNet-V2 outperforms the $HOG + SVM$ model by 3.1% under the first (open-set) protocol. In the case of the more challenging second protocol (heterogeneous open set plus cross-dataset), the MobileNet-V2 model significantly outperforms the $HOG + SVM$ model by a margin of 14.3% in terms of accuracy. Thus the deep learning model is more robust and can generalize better when predicting facial out of sample expressions across datasets and users.

The confusion matrix for the deep learning model and $HOG + SVM$ model accuracy in open-set evaluation on the FACES dataset is shown in Table III. From these results, we can observe:

1) For individual facial expressions, *Anger* emotion classification enjoyed the highest accuracy increase -almost 15%- when switching from the handcrafted features method to the deep learning model.
2) *Happiness* emotion was the next best performing for the deep learning model with more than 4% in accuracy gains when compared to $HOG + SVM$.
3) However, in the case of detecting *Sadness*, there is a 3% drop in accuracy for the deep learning method in comparison to $HOG + SVM$.
4) In both evaluated models, expressions for *Anger* and *Disgust* are the most miss-classified followed by *Anger* and *Sadness*.

Table IV shows the confusion matrix for cross dataset comparison with both the WideNet-V2 model and the $HOG+SVM$ model, trained on the FACES dataset and tested on the CFD dataset. From the results we observe:

1) For individual expression classification, the deep learning model, showed substantial improvement in classifying *Neutral* by 40.85% in accuracy.
2) When it comes to classifying Happiness, the deep learning model miss-classified the expression as Neutral, and in the case of $HOG + SVM$ model, the *Neutral* state is mostly miss-classified as *Happiness*. This show the

978-1-7281-6926-2/20/$31.00 ©2020 IEEE

challenge in differentiating *Happiness* and *Neutral* face expressions in CFD dataset when using the periocular region.

3) Similarly, in the case of *Anger*, both models showed low accuracy. In case of $HOG + SVM$ model, *Anger* is mainly miss-classified as *Disgust*, and with deep learning method as *Neutral*. Thus the narrow ocular ROI may not convey enough information for the aforesaid expressions.

IV. Conclusion

We presented a cross-dataset facial expression recognition from the periocular region. We compared a light deep learning model with a handcrafted feature extraction model. For the deep learning method, we used the MobileNet-V2 model pre-trained on the ImageNet dataset and fine-tuned on emotion expression datasets. For the more traditional handcrafted features method, we used dense HOG features SVM classifiers. Both models have trained on the FACES dataset with $2,052$ samples from 171 Caucasian volunteers with six facial expressions and tested on Chicago face dataset (CFD) with $1,207$ samples from 597 volunteers from different ethnicities. We showed that the MobileNet-V2 deep learning based model can outperform the traditional handcrafted features method by 14.3% in average accuracy. We also conducted open-set evaluation within the same dataset, where identities were divided into non-overlapping training, validation, and testing splits. In this experiment, we show the MobileNet-V2 achieved 3.1% higher accuracy over HOG with SVM, meaning that the advantages of the deep learning model were not as nuanced over a more heterogeneous dataset

As a part of future work, we wish to corporate more extensive and more diverse datasets for training and evaluation under cross-dataset, open set regime, especially for the data-hungry deep learning models.

References

[1] C. E. Izard, "Anxiety: A variable combination of interacting fundamental emotions," *Anxiety: Current trends in theory and research*, vol. 1, pp. 55–106, 1972.

[2] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, p. 124, 1971.

[3] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2014.

[4] I. M. Revina and W. S. Emmanuel, "A survey on human face expression recognition techniques," *Journal of King Saud University-Computer and Information Sciences*, 2018.

[5] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.

[6] W.-L. Chao, J.-J. Ding, and J.-Z. Liu, "Facial expression recognition based on improved local binary pattern and class-regularized locality preserving projection," *Signal Processing*, vol. 117, pp. 1–10, 2015.

[7] Z. Wang and Z. Ying, "Facial expression recognition based on local phase quantization and sparse representation," in *2012 8th International Conference on Natural Computation*. IEEE, 2012, pp. 222–225.

[8] C. F. Liew and T. Yairi, "Facial expression recognition and analysis: a comparison study of feature descriptors," *IPSJ transactions on computer vision and applications*, vol. 7, pp. 104–120, 2015.

[9] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Proceedings Third IEEE international conference on automatic face and gesture recognition*. IEEE, 1998, pp. 200–205.

[10] Y. Liu, J. Wang, and P. Li, "A feature point tracking method basedon the combination of sift algorithm and klt matching algorithm," *Journal of Astronautics*, vol. 32, no. 7, pp. 1618–1625, 2011.

[11] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *2016 IEEE Winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–10.

[12] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2018.

[13] S. Minaee and A. Abdolrashidi, "Deep-emotion: Facial expression recognition using attentional convolutional network," *arXiv preprint arXiv:1902.01019*, 2019.

[14] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial expression analysis under partial occlusion: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 2, pp. 1–49, 2018.

[15] D. H. Lee and A. K. Anderson, "Reading what the mind thinks from how the eye sees," *Psychological science*, vol. 28, no. 4, pp. 494–503, 2017.

[16] F. Alonso-Fernandez, J. Bigun, and C. Englund, "Expression recognition using the periocular region: A feasibility study," in *2018 14th International Conference on Signal-Image Technology & Internet-Based Systems (SITIS)*. IEEE, 2018, pp. 536–541.

[17] 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 *2010 ieee computer society conference on computer vision and pattern recognition-workshops*. IEEE, 2010, pp. 94–101.

[18] E. H. Langholz, "Oculum afficit: Ocular affect recognition," *ArXiv*, vol. abs/1905.09240, 2019.

[19] M. Sandler, A. Howard, M. Zhu, A. Zhmoginov, and L.-C. Chen, "Mobilenetv2: Inverted residuals and linear bottlenecks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 4510–4520.

[20] N. C. Ebner, M. Riediger, and U. Lindenberger, "Faces—a database of facial expressions in young, middle-aged, and older women and men: Development and validation," *Behavior research methods*, vol. 42, no. 1, pp. 351–362, 2010.

[21] D. S. Ma, J. Correll, and B. Wittenbrink, "The chicago face database: A free stimulus set of faces and norming data," *Behavior research methods*, vol. 47, no. 4, pp. 1122–1135, 2015.

[22] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision (IJCV)*, vol. 115, no. 3, pp. 211–252, 2015.

[23] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *CoRR*, vol. abs/1412.6980, 2014.

[24] D. E. King, "Dlib-ml: A machine learning toolkit," *Journal of Machine Learning Research*, vol. 10, pp. 1755–1758, 2009.

978-1-7281-6926-2/20/$31.00 ©2020 IEEE