An Efficient Multi-view Facial Expression Classifier Implementing on Edge Device

Muhamad Dwisnanto Putro, Duy-Linh Nguyen, Adri Priadana, and Kang-Hyun Jo

Department of Electrical, Electronic, and Computer Engineering, University of Ulsan, Ulsan, Korea
dputro@mail.ulsan.ac.kr; ndlinh301@mail.ulsan.ac.kr;
priadana3202@mail.ulsan.ac.kr; acejo@ulsan.ac.kr

Abstract. The robotic technology demands human-robot interaction to implement a real-time facial emotion detector. This system has a role in recognizing the expressions of the user. Therefore, this application is recommended to work quickly to support the robot's capabilities. It helps the robot to analyze the customer's face effectively. However, the previous methods weakly recognize non-frontal faces. It is caused by the facial pose variations only to show partial facial features. This paper proposes a multi-view real-time facial emotion detector based on a lightweight convolutional neural network. It offers a four-stage backbone as an efficient feature extractor that discriminates specific facial components. The convolution with Cross Stage Partial (CSP) approach was employed to reduce computations from convolution operations. The attention module is inserted into the CSP block. These modules also support the detector to work speedily on edge devices. The classification system learns the information about facial features from the KDEF dataset. As a result, facial emotion recognition achieves comparative performance to other methods with an accuracy of 97.10% on the KDEF, 73.95 on the FER-2013, and 84.91% on the RAFDB dataset. The integrated system using a face detector shows that the system obtains a data processing speed of 30 frames per second on the Jetson Nano.

1 Introduction

A robot is required to work automatically and has the capability of perception and action. Perception is the source of information, while the output is an action produced by the robot. Both components cooperate to achieve the goal. Besides, interaction with humans has a social purpose when the robot is implemented in a public area. Human-robot interaction (HRI) has a role in connecting and synchronizing information between robots and users. It implies a closer interaction and demands communication between the both. In addition, they share the workspace in terms of task achievement [19]. Therefore, the misunderstanding of perception will impact the mistake of robot action and incompatibility with the aim. Meanwhile, vision is an essential perceptual attribute to understanding the

environment. Object information is the reference of decisions for the robot to do something. Shape, texture, space, color, and value are the fundamental elements of visual information. It is used as simple knowledge and related to identifying an object.

Robotic vision has been widely implemented to support HRI. A service robot utilizes this technology to recognize user emotions. It is non-verbal communication that is useful for the robot to understand and evaluate the actions. Humans usually show certain expressions on purpose, but they may accidentally occur caused by feelings or emotions. There are six basic human expressions: fear, anger, disgust, surprise, sadness, and happiness[4]. Each emotion presents different feature textures and shapes. It has resulted from one or more movements of muscles in the face. Hence, facial features are the critical element in identifying human emotions. This attention focuses on the facial area [13]. Although gender affects the tendency to represent certain emotions, it has the same facial feature characteristic. The texture identification of facial features is closely related to the success of recognizing human emotions. Besides, this is also influenced by the relationship between facial components in each expression.

Computer vision employs feature extraction to discriminate specific features from the background. Then, it uses a classifier to predict the probability of each category. Several works have used conventional feature extraction [17, 16, 18], but this is not robust for non-frontal faces. This problem does not fully present the essential components of the face. Additionally, rotation-invariant decreases the classifier's performance and causes a classification system to produce high false positives. Convolutional Neural Network (CNN) is an excellent facial feature extraction [24]. It implements a weighted kernel to distinguish the important features of an object. Then, it employs back-propagation to update those weights. This approach delivers high performance for classification tasks. Therefore, several studies have applied it for facial expression recognition work [22, 15, 5]. Recently, various backbone architectures have been presented to distinguish distinctive object features clearly. However, the CNN model requires high GPU usage to work in real-time, while this accelerator is not cheap. On the other hand, computer vision methods are encouraged to be implemented in an edge device such as a Jetson Nano [12]. This device is compatible with sensors and actuators commonly used in robots.

Based on the previously mentioned problems, a real-time facial emotion detector is proposed to recognize multiple poses on basic human facial emotions. The main contributions of this work are as follows:

1. A new facial expression detector is offered to efficiently localize and identify the human facial emotion in different face profiles.
2. A CNN-based light architecture applies Cross Stage Partial (CSP) with an excitation module to reduce the computational operations and parameters.
3. The classification system achieves comparative performance to other methods and performs at a real-time processing speed of 30 FPS on a Jetson Nano.

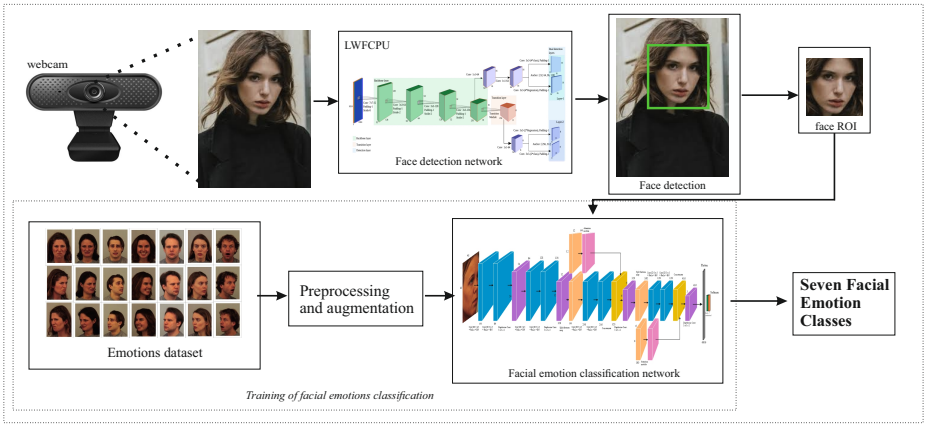

Fig. 1. The overview system of real-time facial emotions detector. It combines face detection and facial expression classification. LWFCPU is used as a face detector to quickly localize medium and large sized faces.

2 Related Works

Several works have applied the CNN approach to classifying facial expressions. Webb et al. [22] have proposed a pre-trained of Deep Convolutional Neural Network (CNN) as a Stacked Convolutional Autoencoder (SCAE) to recognize human emotions that will be implemented in social robots. Transfer learning learns facial features in a greedy layer-wise unsupervised fashion more efficiently. On the other hand, the Multi-model network has obtained higher accuracy for classifying facial expressions in various illuminations and poses [15]. This residual CNN is used to extract specific facial features effectively. Combining 1×1 and 3×3 convolution allows the network to save the computation. In addition, a Squeeze-and-Excitation (SE) Module [7] is applied to the residual block to enhance interest features. Furthermore, Fareed et al. [5] implemented a face localization method at the beginning of the network using Dual Shot Face Detection (DSFD) to overcome the pose invariance. It uses a combination of Linear Discriminant Analysis (LDA) and Adaptive Boosting to re-extract the detected features. Although it obtains high performance, this backbone produces a lot of parameters and expensive computation. It requires a large amount of GPU memory when working in real-time. Multi-view facial expression classifier has been proposed by [1]. It uses a deep convolutional neural network with a transfer learning approach to discriminate the essential features. It applies DenseNet-161 architecture to extract the facial information, so this model tends to run slow when implemented on low-cost devices. Another work [10] uses MobileNet with a convolutional block attention module that can operate fast on cheap devices. However, it is not robust to discriminate the facial features for multi-profile.

3 Proposed Architecture

In this section, a real-time system consists of a face detector and facial expression classification, as shown in Fig. 1. The face detection method uses LWFCPU [14] to generate face ROI (Region of Interest). Furthermore, the facial emotion classification system includes a light backbone with the attention network and a classifier.

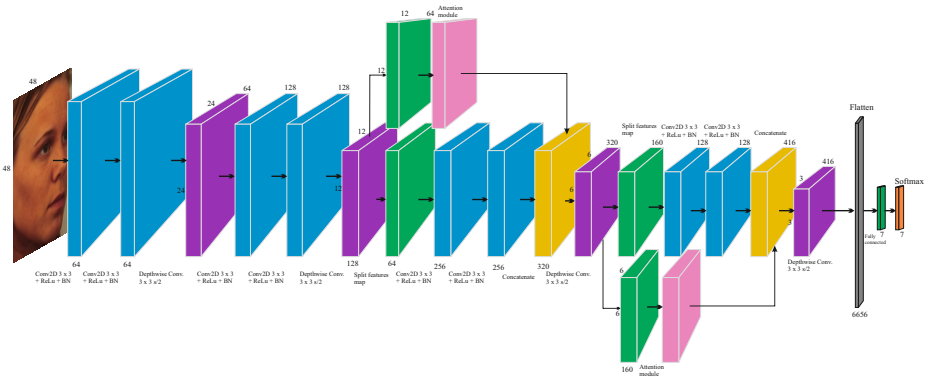

Fig. 2. The proposed architecture of real-time facial expression classification. It uses Cross Stage Partial (CSP) in two stages to reduce the number of operations on the convolutional layer.

3.1 Four-stage light backbone

A CNN-based classification system relies upon the extractor features as an essential module to produce specific features. Each facial expression shows different facial organ information. It means that facial features are critical elements for recognizing each emotion. A four-stage light backbone was introduced using a sparse convolution operation. Fig. 2 shows that this architecture consists of four stages using two 3×3 convolution layers. This block employs an effective filter to find the interest element for identifying the expression.

Furthermore, the proposed architecture applies a Cross Stage Partial (CSP) technique [21] that splits a feature map into two parts with the same number of channels. Then one chunk is transferred and aggregated to the end of the stage. It reduces the computation power of convolution operation without significantly degrading the extractor performance. Reducing the number of channels produces fewer computation costs than the normal process. It also saves the number of parameters. Additionally, the transfer layer avoids losing information caused by the splitting process, which cannot explore these interest elements at the next stage. Therefore, a CSP is only implemented in the third and fourth stages, containing medium and high-level features.

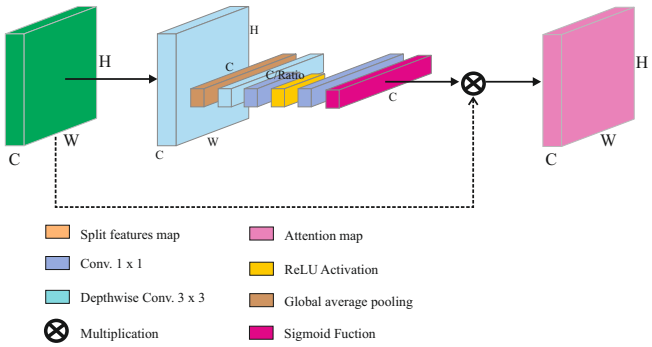

Fig. 3. A depthwise excitation module using weighted vector.

A superficial architecture is weak in discriminating facial features at a high-level frequency. Thus, the attention module enhances specific facial features related to each expression [20]. The proposed architecture develops a depthwise excitation module inserted at each skip connection of the CSP method. It highlights the intensity of the relationship between the facial components in a chunk map. A depthwise convolution is employed as a simple filter that keeps the number of channels of the filter equal to the input, as shown in Fig. 3. Then, Global Average Pooling (GAP) summarizes the intensity of the features as expressed as

$$s_i = W_{d2} \odot GAP(W_{d1} \odot x_i), \tag{1}$$

where $\odot$ is a linear operation of a depthwise convolution applied after and before pooling. It robustly extracted a representation of the essential features s_i. Furthermore, the output of the attention network can be illustrated as

$$At_i = x_i \cdot \sigma(W_{v2} ReLU(W_{v1} s_i)), \tag{2}$$

where σ is the sigmoid function on the sequential operations of the 1×1 convolutional and $ReLU$ activation. Finally, the input features are scaled with a weighted feature representation to update specific features. A depthwise excitation module combines a linear filter and the sequential weighted extraction. It enhances the quality that discriminates useful features and reduces the intensity of trivial features without significantly increasing computation.

3.2 Classifier module

The backbone module generates a 3×3 feature map with 416 channels. The flatten method is applied to reshape tensors into raw vectors. This technique prevents information loss in the classification process. Instead of using the multi fully connected layers, it only uses a dense layer to compress the network parameters. It directly creates a vector with a size that matches the number of emotion

categories. Furthermore, the proposed module applies the Softmax function to generate an output of prediction from the logit score. This activation produces a probability value of each emotion class in the last layer of the neural network. It predicts a multinomial probability distribution in which the sum of all predictions is one.

4 Dataset, Augmentation, and Configuration

4.1 Dataset

The proposed classification system is trained and evaluated on the Karolinska Directed Emotional Faces (KDEF) dataset [2]. This dataset was produced by Karolinska Institutet that consists of 4900 images of human facial expressions. It contains 70 individuals showing seven different emotional faces (neutral, happy, angry, fear, disgusted, sad, and surprised). The subjects are between 20 and 30 years of age. They did not wear earrings, eyeglasses, and makeup in the photo session and did not have beards and mustaches. It also provides five different angles (full left profile, half left profile, straight, half right profile, full right profile) with a 562 $\times$ 762 pixels resolution. On the other hand, the proposed model also evaluates on FER-2013 dataset [6] to comprehensively investigate its performance. This dataset provides 35,887 pictures with 48 $\times$ 48 pixels. The grayscale image covers seven emotions: anger, neutral, sad, fear, happy, surprise, and disgust. In addition, this database is a challenging dataset that inserts a few invalid labels. Additionally, it also utilizes the RAFDB [9] dataset to examine the proposed model. This dataset provides 30,000 facial images annotated. Our model uses basic emotion that contains 12,271 images for training and 3,068 images for testing.

4.2 Preprocessing and augmentation

A face detector [14] is applied to the image dataset to generate the ROI of the face. It encourages the classification model to focus on facial areas. The training and evaluation process uses the RGB images with 48 $\times$ 48, resized from facial ROI. To expand the training dataset, this applies the augmentation method. Additionally, this technique also improves the performance and ability of the real-time detector. The first stage is manipulating various lighting using random contrast, brightness, saturation, and hue. Then, it implements multiple rotations to enrich the variety of facial poses. The last process is to apply a horizontal flip to the entire previous augmented image. In contrast, the augmentation technique is not utilized in the FER-2013 and RAFDB datasets.

4.3 Training configuration

The training of the classification model uses several configurations and parameters. This setting helps optimize the training process. Categorical cross-entropy

is used to calculate the loss of the prediction into the ground truth. Meanwhile, Adam (Adaptive Moment Estimation) is utilized to optimize this process with an epsilon of 10^{-7}. The KDEF dataset starts with the 10^{-4} learning rate. It then will be updated by multiplying 0.75 when the accuracy does not improve in 20 epochs. The proposed model was conducted in the Keras framework. The training uses a batch size of 128 with epochs of 50 on 10-fold. It uses K-fold cross-validation to split and evaluate the model. On the other hand, our model trained with on FER-2013 dataset that provides 3,589 images. It uses a 10^{-4} learning rate and 64 batch sizes in the 500 epochs. Besides, It sets 200 epochs, 32 batch sizes, and 10^{-4} learning rate for training configuration on the RAFDB dataset.

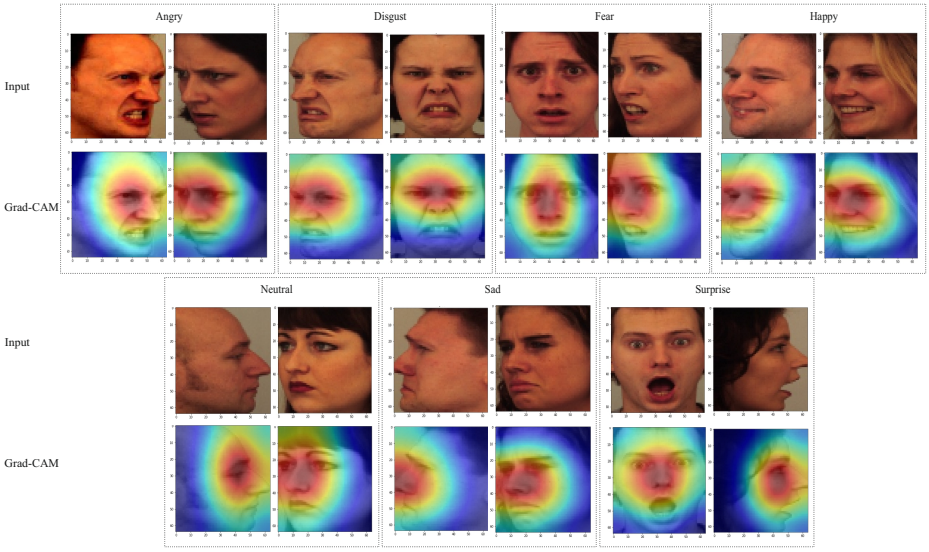

Fig. 4. Result of heat map attention for seven classes of emotions using GRAD-CAM approach.

5 Experiments and Results

5.1 Ablative study

The proposed architecture consists of several modules that corporates to improve performance and efficiency. The ablation study is conducted to examine the effect of each module. The offered modules are applied one by one to analyze the strength, as shown in Table 1. Firstly, the Fours-stages backbone is proposed as a shallow-layered feature extractor. This backbone obtains an accuracy of 72.50% and generates 2.36M parameters. Secondly, to reduce the training parameters by

Table 1. Ablative result of proposed architecture on the FER-2013 dataset

Modules	**Proposed model**		
Four-stage backbone	√	√	√
Cross stage partial		√	√
Depthwise excitation			√
Parameters	2,353,223	1,996,903	**2,006,759**
Accuracy(%)	72.50	71.55	**73.95**

0.35M, increasing the detector's efficiency, it applies the CSP approach without significantly decreasing the accuracy by 0.95%. Finally, the depthwise excitation module increases the accuracy by 2.40% without adding many parameters.

The attention module improves the quality of input features by strengthening the intensity of essential elements. Additionally, it also captures the relationship between these components that are related to each expression. Fig. 4 shows that the proposed module concentrates and focuses on specific areas and organs of the face. The heat map indicates useful facial features to classify facial emotions. It can precisely localize the facial area of interest for the non-frontal face. Moreover, the proposed model pays attention to the eyes, eyebrows, nose, and cheeks. These elements are related to each other to generate certain emotions.

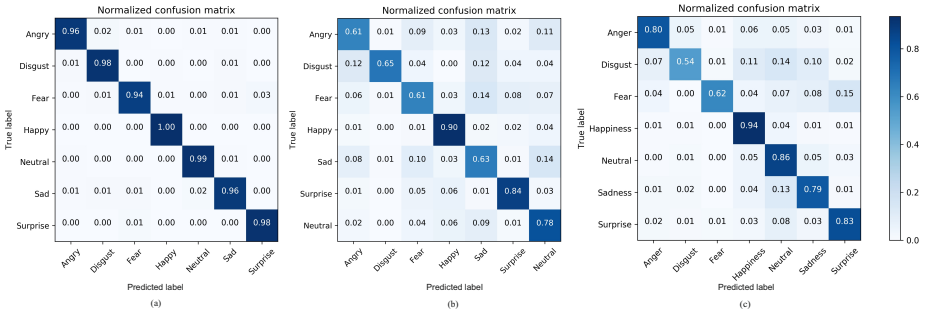

Fig. 5. Confusion matrix of evaluation at each emotion categories on KDEF (a), FER-2013 (b) and RAFDB (c) dataset.

5.2 Comparison on datasets

This paper proposes the light backbone to distinguish essential features that can work in real-time. It uses a shallow layered convolution and the Cross Partial Stage to generate 1,996,903 parameters. The proposed classification system is trained and evaluated on the KDEF dataset. The evaluation results are also compared with previous methods. Table 2 shows that the proposed architecture

Table 2. Evaluation of proposed architecture compared to other methods on KDEF, FER-2013 and RAFDB dataset

KDEF dataset	
Method	**Accuracy**
O-FER[3]	91.42
CCFN[23]	91.60
Multi-Xception	94.29
Resnet-19	94.49
Multi-C-Xception	94.63
DFSD-LDA-AdaBoost	95.06
Akhand et al[1]	96.51
Proposed	97.10
FER-2013 dataset	
Multi-scale CNN	72.82
SNNs	73.00
Single MLCNN[11]	73.03
Ensemble MLCNNs[11]	74.09
AM-Net[8]	75.82
Proposed	73.95
RAFDB dataset	
MobileNetV1	81.62
PG-CNN	82.27
DLP-CNN	84.22
A-MobileNet[10]	84.49
Proposed	84.91

achieves an accuracy of 97.10%. This result is superior to all facial expression methods that have been present in the KDEF dataset. The proposed model outperforms 1.59% of the Multi-fusion model incorporating a CNN-based residual and a Squeeze-and-Excitation (SE) module. As shown in Fig. 5 (a), the prediction for each category is analyzed in the confusion matrix. The dark color indicates high accuracy obtained by each matrix element, and the bright color is vice versa. The proposed model performs best when it predicts "Happy." This emotion has a unique facial shape compared to other expressions. Meanwhile, "Fear" obtained the lowest score. Some instances are predicted "surprises" because both emotions show similar shapes and textures. Our model is also examined in realistic profile variation scenarios. As shown in Table 3, each profile is investigated that evaluated on the KDEF dataset. The proposed model is powerful in recognizing the center position that achieves the highest accuracy of 98.57%. In contrast, the full left pose obtains low accuracy of 95.59%.

Furthermore, the proposed model examines difficult challenges using the FER-2013 dataset. It reached an accuracy of 73.95%, which is weaker than Ensemble MLCNN[11] and AM-NET[8]. However, it outperforms SSNs and single MLCNN[11], with a 1% difference. Our model is robust in recognizing the

Table 3. Evaluation of multi-profile facial expression

Face pose	**Accuracy on KDEF(%)**
Full left	95.59
Full right	95.72
Half left	98.23
Half right	98.27
Center	98.57

Table 4. Runtime efficiency compared to competitors on Jetson Nano

Method	**Paremeter**	**Accuracy(%)**			**Speed of**
		KDEF	**FER-2013**	**RAFDB**	**Integrated (FPS)**
Multi-model fusion[15]	1,206,279	93.42	-	-	26.38
AM-NET[8]	24,904,204	-	75.82	-	5.71
MLCNN[11]	20,787,783	-	73.03	-	10.89
Ensemble MLCNN[11]	92,825,543	-	74.09	-	Out of memory
Akhand et al[1]	28,907,943	96.51	-	-	Out of memory
A-MobileNet[10]	3,321,513	-	**-**	84.49	20.35
Proposed detector	**2,006,759**	**97.10**	**73.95**	**84.91**	29.58

"Happy" category, as illustrated in Fig. 5 (b). in contrast, it not powerful to classify "Angry" and "Fear" on this dataset. The proposed model also examines the performance using the RAFDB dataset, which shows that it is superior to other models. It higher 0.42% than [10], as shown in Table 3. Although the proposed model obtains low accuracy in recognizing "Disgust" emotion, It achieves accurate prediction in the "Happiness" category.

5.3 Real-time application

The practical application requires a vision-based detector to work in real-time. In addition, robotic technology implements it on edge devices, which are compatible with sensors and actuators. Hence, a real-time face emotion detector on the Jetson Nano with input from a webcam. It compares the proposed detector's speeds to a competitor integrated with a face detector [14], as shown in Fig. 1. Face detection produces facial ROI to avoid perturbation of background features. Table IV shows that the Multi-model fusion has a small number of training parameters. However, the proposed detector achieves a more accurate performance on the classification system and requires a data processing speed of 29.58 FPS. Although AM-NET[8] and Ensemble MLCNN[11] achieve high accuracy, our model is faster than the competitors. The proposed model is even faster than A-MobileNet[10] and Akhand et al[1]. It proves that the proposed detector more efficiently works on an edge device. The two-stage detector sequentially employs face detection and a classification system to predict facial areas and classify them. Therefore, face detection is a mandatory process for generating

Fig. 6. Qualitative results in real-time application with multiple faces.

facial patches. Then, the classification network estimates the emotion category. The effectiveness of the detector performance in real applications is shown in Fig. 6. It robustly detects and classifies the expression of multiple facial poses. This system is feasible to be implemented in robots to support human-robot interaction.

6 Conclusions

This paper proposes a real-time facial emotion detector to predict seven classes of multi-profile facial expressions implemented on an edge device. The integrated system can support the human-robot interaction system. The proposed architecture consists of a four-stage backbone to efficiently extract the specific features and a depthwise excitation module to increase the intensity of the useful features. The CSP approach improves the model's efficiency without significantly reducing the detector performance. In order to build a robust model implemented in real applications, a classification system is trained and evaluated on the KDEF and the FER-2013 dataset that provides multi-profile face instances. As a result, the proposed model achieves high accuracy, competitive with the previous methods. In addition, the integrated system achieves a data processing speed of

29.57 FPS when working on a Jetson Nano. Future work will explore the margin loss to improve accuracy by balancing true and false prediction losses.

Acknowledgment

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the government (MSIT). (No.2020R1A2C200897212).

References

1. Akhand, M.A.H., Roy, S., Siddique, N., Kamal, M.A.S., Shimamura, T.: Facial emotion recognition using transfer learning in the deep cnn. Electronics 10(9) (2021)
2. Calvo, M., Lundqvist, D.: Facial expressions of emotion (kdef): Identification under different display-duration conditions. In: Behavior Research Methods. vol. 40, p. 109–115 (1998), http://www.kdef.se/
3. Dong, J., Zhang, L., Chen, Y., Jiang, W.: Occlusion expression recognition based on non-convex low-rank double dictionaries and occlusion error model. Signal Processing: Image Communication 76, 81–88 (2019)
4. Ekman, P.: Facial expressions of emotion: New findings, new questions. Psychological Science 3(1), 34–38 (1992)
5. Fareed, K., Sultan, F., Khan, K., Mahmood, Z.: A robust face recognition method for expression and pose variant images. In: 2020 14th International Conference on Open Source Systems and Technologies (ICOSST). pp. 1–6 (2020)
6. Goodfellow, I.J., Erhan, D., Luc Carrier, P., Courville, A., Mirza, M., Hamner, B., Cukierski, W., Tang, Y., Thaler, D., Lee, D.H., Zhou, Y., Ramaiah, C., Feng, F., Li, R., Wang, X., Athanasakis, D., Shawe-Taylor, J., Milakov, M., Park, J., Ionescu, R., Popescu, M., Grozea, C., Bergstra, J., Xie, J., Romaszko, L., Xu, B., Chuang, Z., Bengio, Y.: Challenges in representation learning: A report on three machine learning contests. Neural Networks 64, 59 – 63 (2015), special Issue on "Deep Learning of Representations"
7. Hu, J., Shen, L., Sun, G.: Squeeze-and-excitation networks. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 7132–7141 (2018)
8. Li, J., Jin, K., Zhou, D., Kubota, N., Ju, Z.: Attention mechanism-based cnn for facial expression recognition. Neurocomputing 411, 340 – 350 (2020)
9. Li, S., Deng, W.: Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Transactions on Image Processing 28(1), 356–370 (2019)
10. Nan, Y., Ju, J., Hua, Q., Zhang, H., Wang, B.: A-mobilenet: An approach of facial expression recognition. Alexandria Engineering Journal 61(6), 4435–4444 (2022), https://www.sciencedirect.com/science/article/pii/S1110016821006682
11. Nguyen, H.D., Kim, S.H., Lee, G.S., Yang, H.J., Na, I.S., Kim, S.H.: Facial expression recognition using a temporal ensemble of multi-level convolutional neural networks. IEEE Transactions on Affective Computing 13(1), 226–237 (2022)
12. Pathak, R., Singh, Y.: Real time baby facial expression recognition using deep learning and iot edge computing. In: 2020 5th International Conference on Computing, Communication and Security (ICCCS). pp. 1–6 (2020)

13. Putro, M.D., Jo, K.: Real-time face tracking for human-robot interaction. In: Proceedings of the International Conference on Information and Communication Technology Robotics (ICT-ROBOT). pp. 1–4 (Sep 2018)
14. Putro, M.D., Nguyen, D., Jo, K.: Lightweight convolutional neural network for real-time face detector on cpu supporting interaction of service robot. In: 2020 13th International Conference on Human System Interaction (HSI). pp. 94–99 (2020)
15. Qi, A., Wei, J., Bai, B.: Research on deep learning expression recognition algorithm based on multi-model fusion. In: 2019 International Conference on Machine Learning, Big Data and Business Intelligence (MLBDBI). pp. 288–291 (2019)
16. Rao, Q., Qu, X., Mao, Q., Zhan, Y.: Multi-pose facial expression recognition based on surf boosting. In: 2015 International Conference on Affective Computing and Intelligent Interaction (ACII). pp. 630–635 (2015)
17. Rujirakul, K., So-In, C.: Histogram equalized deep pca with elm classification for expressive face recognition. In: 2018 International Workshop on Advanced Image Technology (IWAIT). pp. 1–4 (2018)
18. Santra, B., Mukherjee, D.P.: Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression. In: 2016 IEEE International Conference on Image Processing (ICIP). pp. 624–628 (2016)
19. Sirithunge, C., Ravindu, H.M., Bandara, T., Buddhika, A.G., Jayasekara, P., Chandima, D.P.: Situation awareness for proactive robots in hri. In: 2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). pp. 7813–7820 (2019)
20. Sun, W., Zhao, H., Jin, Z.: A visual attention based roi detection method for facial expression recognition. Neurocomputing 296, 12 – 22 (2018), http://www.sciencedirect.com/science/article/pii/S0925231218303266
21. Wang, C., Mark Liao, H., Wu, Y., Chen, P., Hsieh, J., Yeh, I.: Cspnet: A new backbone that can enhance learning capability of cnn. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 1571–1580 (2020)
22. Webb, N., Ruiz-Garcia, A., Elshaw, M., Palade, V.: Emotion recognition from face images in an unconstrained environment for usage on social robots. In: 2020 International Joint Conference on Neural Networks (IJCNN). pp. 1–8 (2020)
23. Ye, Y., Zhang, X., Lin, Y., Wang, H.: Facial expression recognition via region-based convolutional fusion network. Journal of Visual Communication and Image Representation 62, 1–11 (2019)
24. Zeiler, M.D., Fergus, R.: Visualizing and understanding convolutional networks. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 818–833. Springer International Publishing, Cham (2014)