A Real-time Multi-view Face Emotion Detector on Edge Device

Muhamad Dwisnanto Putro[1], Duy-Linh Nguyen[2], and Kang-Hyun Jo[3]

***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 capabilities of the robot. It helps the robot to effectively analyze the customer's face. However, the previous methods weakly recognize non-frontal faces. It is caused by the facial pose variations only to show partial facial features. In this paper, a multi-view real-time facial emotion detector is developed based on a lightweight convolutional neural network. A four-stage backbone is introduced as an efficient feature extractor that discriminates specific facial components. Cross Stage Partial (CSP) and Depthwise block were 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 of facial features from the KDEF dataset. As a result, facial emotion recognition achieves comparative performance to other methods with an accuracy of 94.54%. The integrated system using a face detector shows that the system obtains a data processing speed of 35 frames per second on the Jetson Nano.**

I. INTRODUCTION

A robot is required to work automatically and has the capability of perception and action. Perception is the source of information, while action is the output 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 the 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. They share the workspace and the goal in terms of task achievement [1]. 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 understand the environment. Information of object is the reference of decisions for the robot to do something. Shape, texture, space, color, and value are the basic 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 previous actions. Humans usually show certain expressions on purpose, but they may accidentally occur caused by feelings or emotions. Anger, fear, disgust, happiness, sadness, and surprise are six basic human expressions [2]. Each emotion presents a different feature texture and shape. It has resulted from one or more movements of muscles in the face. Therefore, facial features are the critical element to identifying human emotions so that the focussed attention only on the facial area [3]. Although gender affects the tendency to express certain emotions, it still creates the same facial feature actions. 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 at each emotion.

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 [4], [5], [6], 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 [7]. It implements a weighted kernel to distinguish the essential 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 [8], [9], [10]. 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 [11]. 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) CNN-based light architecture applies Cross Stage Partial (CSP) and depthwise convolution block, which emphasizes reduction of computational operations and parameters.
2) The classification system achieves comparative performance to other methods and performs in real-time by 30 FPS on a Jetson Nano.

*This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT). (No. 2020R1A2C2008972)

[1,2,3]Department of Electrical, Electronics and Computer Engineering, University of Ulsan, Ulsan, Korea [1]dputro@mail.ulsan.ac.kr, [2]ndlinh301@mail.ulsan.ac.kr, [3]acejo@ulsan.ac.kr

Manuscript 555 submitted to 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Received March 3, 2021.

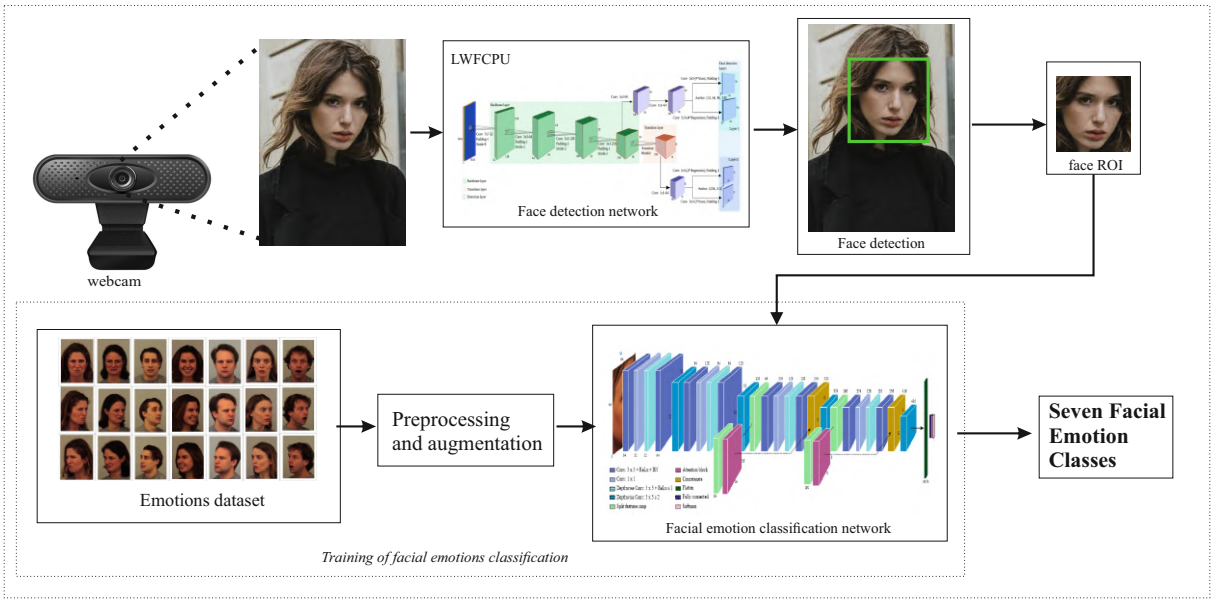

Fig. 1. The overview system of real-time facial emotions detector. It combines face detection and facial expression classification. LWFCPU is used as a face detector to quickly localize medium and large sized faces.

II. RELATED WORKS

Several works have been applied the CNN approach to classifying facial expressions. Webb et al. [8] 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 helps models to learn facial features in a greedy layer-wise unsupervised fashion more efficiently. Then, it fuses convolutional and fully connected layers to encode facial expression images in a features vector. On the other hand, the Multi-model network has obtained higher accuracy for classifying facial expressions in various illumination and attitude angles [9]. This residual-CNN is used to extract specific facial features effectively. The combination of 1 × 1 and 3 × 3 convolution allows the network to save the computation. In addition, a Squeeze-and-Excitation (SE) Module [12] is applied to the residual block to enhance useful features. Furthermore, Fareed et al. [10] implemented a face localization method at the beginning of the network using Dual Shot Face Detection (DSFD) to overcome the pose invariance. Then, it uses a combination of Linear Discriminant Analysis (LDA) and Adaptive Boosting to re-extract the detected features. At the end of the network, the classic nearest center classifier is applied to classify facial emotions classes. 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.

III. PROPOSED ARCHITECTURE

In this section, the overall real-time system consists of a face detector and emotion classification, as shown in Fig. 1. The face detection method uses LWFCPU [13] to generate face ROI (Region of Interest). Furthermore, the facial emotion classification system consists of a light backbone with the attention network and a classifier.

A. *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, which is means that facial features are critical elements to recognize 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 x 3 convolution layers, which flanks a separable depthwise convolution with a smaller number of channels [14]. This formation generates fewer computations and parameters compared to sequentially three times 3 x 3 convolutions. Then, In order to reduce the feature map size to the classification vector, it uses the depthwise convolution with strides of two that is more robust than the pooling methods.

Furthermore, the proposed architecture applies a Cross Stage Partial (CSP) technique [15] that splits a feature map become 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. The reduction in the number of channels produces fewer computation costs than the normal process. The kernel operates with a smaller number of channels, and it also saves the number of parameters. Additionally, the transfer layer avoids losing information caused by the splitting process, which nevertheless explores these interest features at the next stage. Therefore, CSP is only implemented in stages three and four, containing mid-level and high-level features.

Manuscript 555 submitted to 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Received March 3, 2021.

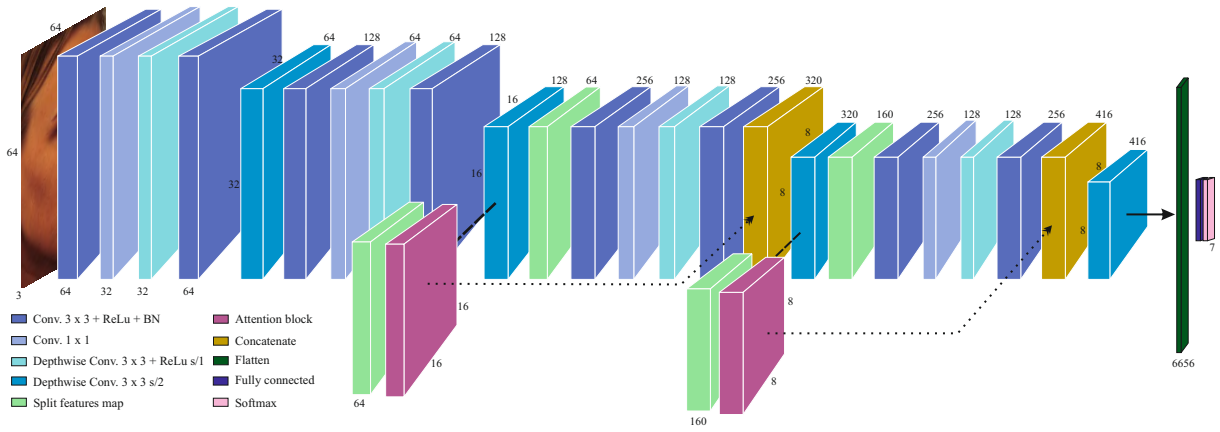

Fig. 2. Proposed architecture of real-time face emotions classification. It uses Cross Stage Partial (CSP) on two stages to reduce the number of operations on the convolutional layer.

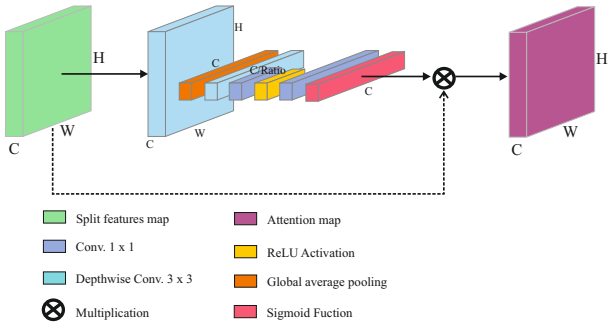

Fig. 3. A depthwise excitation module using weighted vector.

A superficial architecture is weak to discriminate important facial features clearly. Thus, the attention module enhances specific facial features related to each expression [16]. 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 x 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 simple filter and the sequential weighted extraction. It enhances the quality to discriminate useful features and dims the intensity of trivial features without significantly increasing computation.

B. Classifier module

The backbone module generates an 8 x 8 feature map with 416 channels. The Flatten method is applied to reshape tensors into raw vectors. This technique is useful for simplifying the classification process and preventing information loss. 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 generates a probabilities 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.

IV. DATASET, AUGMENTATION, AND CONFIGURATION

A. Dataset

The proposed classification system is trained and evaluated on the Karolinska Directed Emotional Faces (KDEF) dataset [17]. This dataset was produced by Karolinska Institutet that consists of 4900 images of human facial expressions. It contains 70 individuals showing seven different emotional expressions (neutral, happy, angry, fear, disgusted, sad, and surprised). The subjects are between 20 and 30 years of age. In the photo session, they did not wear earrings, eyeglasses, and make-up 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 resolution of 562 x 762 pixels.

B. Preprocessing and augmentation

Face detector [13] is applied to the images 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 64 x 64, which is resized from facial ROI. In order to expand the training

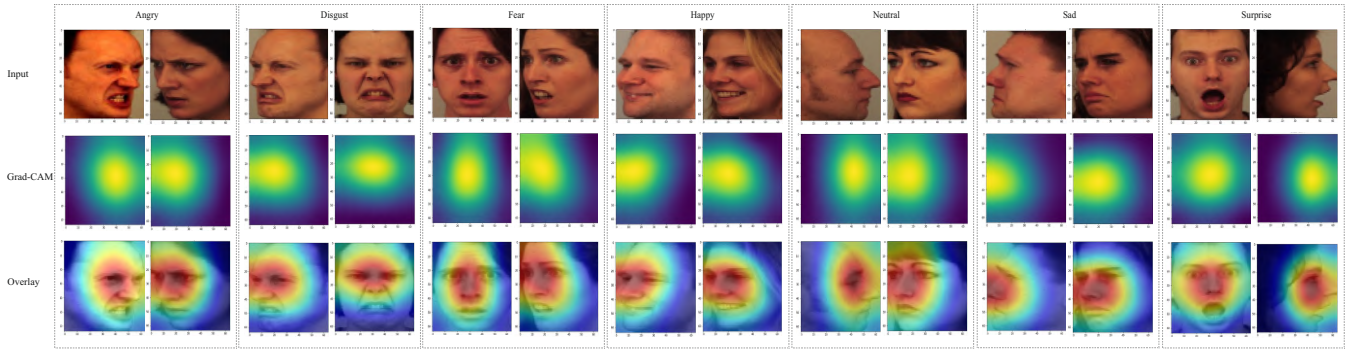

Fig. 4. Result of attention heat map for seven classes of emotions using GRAD-CAM approach.

dataset, this applies the augmentation method. Additionally, this technique is also to improve the performance and ability of the real-time detector. The first stage is to manipulate a variety of lighting using low-high contrast and brightness. Then, it implements multiple rotations (0 - 50 degrees) to enrich the variety of facial poses. The last process is to apply a horizontal flip to the entire previous augmented image. As a result, this augmentation method produces 98,000 RGB images.

C. Training configuration

The training of the classification model uses several configurations and parameters. This setting is useful for optimizing 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 as an optimizer with an epsilon of 10^{-7}. The training uses a batch size of 128 with total epochs of 300. In order to optimize the training phase, it starts with the 10^{-4} learning rate. It will be updated by multiplying 0.75 when the accuracy does not improve in 25 epochs. The proposed model was conducted in the Keras framework.

V. EXPERIMENTS AND RESULTS

A. 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 proposed modules are applied one by one to analyze the strength, as shown in Table I. Firstly, the Fours-stages backbone is proposed as a shallow-layered feature extractor. This backbone obtains an accuracy of 94.12(%) and generates 2.36M parameters. Secondly, a depthwise separable is applied between two 3 x 3 convolutions at each stage. It uses a smaller number of channels than input features, thereby reducing parameters and improving accuracy slightly. This experiment obtained the accuracy and number of parameters of 94.24(%) and 1.77M, respectively. Thirdly, in order to reduce the training parameters by 0.35M that is increasing the detector's efficiency, it applies the CSP approach without significantly reducing the accuracy by 0.2(%). Finally, the depthwise excitation module was used to increase the accuracy by 0.85(%) without adding many parameters.

TABLE I

ABLATIVE RESULT OF PROPOSED ARCHITECTURE ON KDEF DATASET

Modules	**Proposed model**			
Four-stage backbone	√	√	√	√
Depthwise separable	x	√	√	√
Cross stage partial	x	x	√	√
Depthwise excitation	x	x	x	√
Parameters	2.36	1.77	1.42	**1.43**
Accuracy(%)	94.12	94.24	94.04	**94.89**

TABLE II

EVALUATION OF PROPOSED BACKBONE COMPARED TO MOBILENET MODEL ON KDEF DATASET.

Methods	**Parameters**	**Accuracy**(%)
MobileNet	3,236,039	85.13
MobileNetV2	2,266,951	84.72
MobileNetV3	5,127,839	83.79
Light backbone	**1,424,807**	**94.04**

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 shows that red areas get more attention than other colors, determining useful facial features to classify facial emotions. Even it can precisely localize the facial area of interest for the non-frontal face. Various expressions, facial poses, and gender prove that the proposed model pays attention to the eyes, eyebrows, nose, and cheeks. These elements are related to each other to generate certain emotions. Although the superficial model ignores the lips, mouth, and chin as important information, it achieves decent performance for classifying facial emotions in real-time using a small accelerator device.

B. Result of comparison

In this paper, the light backbone is proposed to distinguish essential features that can work in real-time. It uses a shallow layered convolution, the depthwise bottleneck, and the Cross Partial Stage to generate 1,416,103 parameters. TABLE II

TABLE III

EVALUATION OF PROPOSED ARCHITECTURE COMPARED TO OTHER METHODS ON KDEF DATASET.

Methods	Accuracy (%)
SRC	89.52
GSRC	89.90
CRC	90.24
PCRC	90.71
RCFN	90.73
RCFN(CPL)	91.11
O-FER [18]	91.42
CCFN [19]	91.60
Multi-Model fusion [9]	93.42
Proposed model	**94.88**

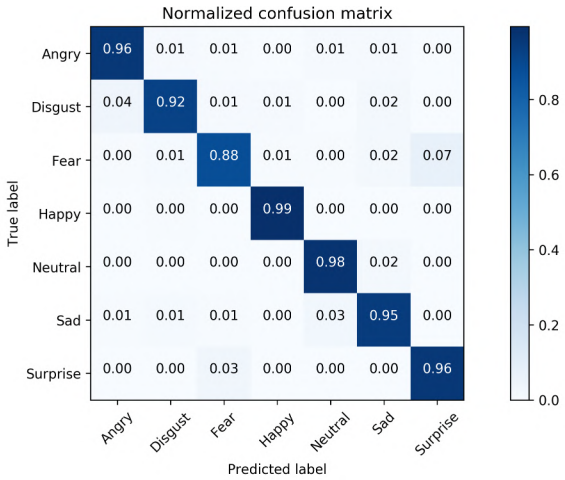

Fig. 5. Confusion matrix of evaluation at each emotion categories.

compares the backbone performance with a baseline benchmark that can be implemented on a mobile device. As a result, the proposed backbone outperforms MobileNet, MobileNetV2, and MobileNetV3. Besides that, it also produces a fewer number of parameters from the models.

The proposed classification system is trained and evaluated on the KDEF dataset. The evaluation results were also compared with previous methods. TABLE III shows that the proposed architecture achieves an accuracy of 94.88%. This result is superior to all facial expression methods that have been present in the KDEF dataset. The proposed model outperforms 1.46% from the Multi-fusion model incorporating a CNN-based residual and a Squeeze-and-Excitation (SE) module. Furthermore, the prediction for each category is analyzed in the confusion matrix, as shown in Fig. 5. 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 a similar shape and texture.

C. 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 that 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, which is integrated with a face detector [13], as shown in Fig. 1. Face detection produces facial ROI to avoid perturbation of background features. TABLE IV shows that the Multi-model fusion produces a smaller number of training parameters. However, the proposed detector achieves a more accurate performance on the classification system and outperforms the data processing speed by 30.35 FPS. It proves that the proposed detector more efficiently works on an edge device.

TABLE IV

RUNTIME EFFICIENCY COMPARED TO A COMPETITOR ON JETSON NANO.

Methods	Parameters	Acc(%)	Speed of integrated system(FPS)
Multi-model fusion [9]	1,206,279	93.42	28.98
Proposed detector	**1,424,807**	**94.9**	**30.35**

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 facial patches. Then, the classification network predicts the emotion category. The effectiveness of the detector performance in real applications is shown in Fig. 6 (a). It robustly detects and classifies the expression of multiple facial poses. Besides, Fig 6 (b) shows that the detector can work effectively on multiple people. This system is feasible to be implemented in robots to support human-robot interaction. Service robots can use it to help their task. This robot tends to interact closely with the user. Therefore, the proposed detector has robust performance when predicting medium and large faces.

VI. CONCLUSIONS

In this paper, a real-time face emotion detector is proposed 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 depthwise convolution blocks and the CSP approach improve the model's efficiency without significantly reducing the detector performance. In order to build a robust model that is implemented in real applications, a classification system is trained and evaluated on the KDEF dataset that provides multi-profile face instances. As a result, the proposed model achieves an accuracy of 94.88%, which outperforms the previous methods. In addition, the integrated system achieves a data processing speed of 30.35 FPS when working on a Jetson Nano. Future work will explore the margin loss to improve accuracy by balancing true and false prediction losses.

Manuscript 555 submitted to 2021 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS). Received March 3, 2021.

Fig. 6. Qualitative results in real-time application with single (a) and multiple people (b).

REFERENCES

[1] C. Sirithunge, H. M. Ravindu, T. Bandara, A. G. Buddhika, P. Jayasekara, and D. P. Chandima, "Situation awareness for proactive robots in hri," in *2019 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, 2019, pp. 7813–7820.

[2] P. Ekman, "Facial expressions of emotion: New findings, new questions," *Psychological Science*, vol. 3, no. 1, pp. 34–38, 1992.

[3] M. D. Putro and K. Jo, "Real-time face tracking for human-robot interaction," in *Proceedings of the International Conference on Information and Communication Technology Robotics (ICT-ROBOT)*, Sep. 2018, pp. 1–4.

[4] K. Rujirakul and C. So-In, "Histogram equalized deep pca with elm classification for expressive face recognition," in *2018 International Workshop on Advanced Image Technology (IWAIT)*, 2018, pp. 1–4.

[5] Q. Rao, X. Qu, Q. Mao, and Y. Zhan, "Multi-pose facial expression recognition based on surf boosting," in *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*, 2015, pp. 630–635.

[6] B. Santra and D. P. Mukherjee, "Local saliency-inspired binary patterns for automatic recognition of multi-view facial expression," in *2016 IEEE International Conference on Image Processing (ICIP)*, 2016, pp. 624–628.

[7] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *Computer Vision – ECCV 2014*, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 818–833.

[8] N. Webb, A. Ruiz-Garcia, M. Elshaw, and V. Palade, "Emotion recognition from face images in an unconstrained environment for usage on social robots," in *2020 International Joint Conference on Neural Networks (IJCNN)*, 2020, pp. 1–8.

[9] A. Qi, J. Wei, and B. Bai, "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)*, 2019, pp. 288–291.

[10] K. Fareed, F. Sultan, K. Khan, and Z. Mahmood, "A robust face recognition method for expression and pose variant images," in *2020 14th International Conference on Open Source Systems and Technologies (ICOSST)*, 2020, pp. 1–6.

[11] R. Pathak and Y. Singh, "Real time baby facial expression recognition using deep learning and iot edge computing," in *2020 5th International Conference on Computing, Communication and Security (ICCCS)*, 2020, pp. 1–6.

[12] J. Hu, L. Shen, and G. Sun, "Squeeze-and-excitation networks," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 7132–7141.

[13] M. D. Putro, D. Nguyen, and K. Jo, "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)*, 2020, pp. 94–99.

[14] V. Hoang, V. Hoang, and K. Jo, "Realtime multi-person pose estimation with rcnn and depthwise separable convolution," in *2020 RIVF International Conference on Computing and Communication Technologies (RIVF)*, 2020, pp. 1–5.

[15] C. Wang, H. Mark Liao, Y. Wu, P. Chen, J. Hsieh, and I. Yeh, "Cspnet: A new backbone that can enhance learning capability of cnn," in *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2020, pp. 1571–1580.

[16] W. Sun, H. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12 – 22, 2018. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0925231218303266

[17] M. Calvo and D. Lundqvist, "Facial expressions of emotion (kdef): Identification under different display-duration conditions," in *Behavior Research Methods*, vol. 40, no. 2008, 1998, p. 109–115. [Online]. Available: http://www.kdef.se/

[18] J. Dong, L. Zhang, Y. Chen, and W. Jiang, "Occlusion expression recognition based on non-convex low-rank double dictionaries and occlusion error model," *Signal Processing: Image Communication*, vol. 76, pp. 81–88, 2019.

[19] Y. Ye, X. Zhang, Y. Lin, and H. Wang, "Facial expression recognition via region-based convolutional fusion network," *Journal of Visual Communication and Image Representation*, vol. 62, pp. 1–11, 2019.