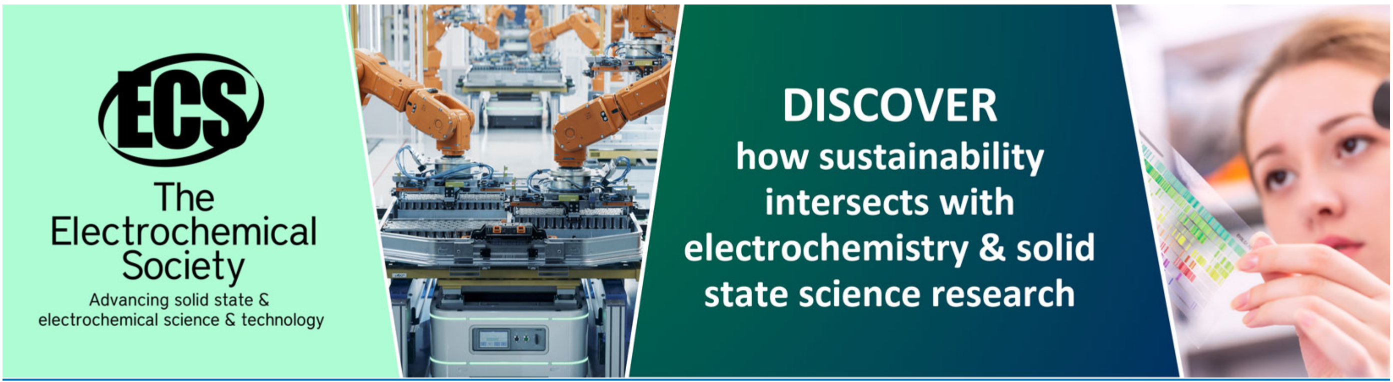
Journal of Physics: Conference Series

LRERNet: Low Resolution Expression Recognition Network

To cite this article: Yipeng Ma *et al* 2020 *J. Phys.: Conf. Ser.* **1621** 012086



ECS
The Electrochemical Society
Advancing solid state & electrochemical science & technology
DISCOVER
how sustainability intersects with electrochemistry & solid state science research

LRERNet: Low Resolution Expression Recognition Network

Yipeng Ma, Feifei Tong and Yuehai Wang

College of Information Science and Electronic Engineering, Zhejiang University,
Hangzhou, China
Email: mayipeng@zju.edu.cn; 21760211@zju.edu.cn; wangyuehai@zju.edu.cn

Abstract. Facial expression recognition (FER) is a meaningful but challenging research direction. An important reason is that the image resolution used for FER is usually low. At present, there is no specially designed model for the challenge of low resolution FER. We propose a low resolution expression recognition network with front and back end structure using specially designed dilated convolution groups. In addition, for the need of pension agency, we establish a dataset for studying FER of the old in Asian countries and we named it AOPFE. We evaluate our method on three standard datasets (RAF, SFEW and CK+) and AOPFE. In our experiments, the method achieved good results in these datasets especially on RAF dataset.

1. Introduction

Facial expression recognition (FER) is a meaningful and challenging research direction in computer vision field, which aims at analysing and identifying human expression states. With the continuous improvement of hardware computing power and the rapid development of AI technology, FER algorithms are also constantly improving. FRE algorithms are divided into symbol-based algorithms and feature-based algorithms according to characteristics. Feature-based algorithms are classified as traditional manual features-based algorithms and deep learning algorithms.

Symbol-based algorithms, represented by FACS [1], divide faces into several action units to analyse the categories of facial expressions through features combination. The method based on traditional manual features first extracts manual features, such as LBP [2], BOW [3], HOG [4] and SIFT [5] and then uses SVM, NN and other classifiers for FER [6].

Deep learning features-based method is the mainstream method of FER, and has achieved good results in laboratory environments [7-11]. But the recognition accuracy of these methods in real scenes is still not ideal. Deeper neural network model can usually bring better recognition effect. However, the resolution of image used for expression recognition is generally low. With the increase of the depth of the neural network and pooling times, it is not possible to effectively extract deep features while increasing the field of receptivity, which makes the recognition performance terrible. For this question, there is no relevant solution now. In this paper, a novel front-back end CNN model is proposed. The front-end network is a typical CNN model. The back-end network introduces a specially designed combination of dilated convolutions, which can fully extract deep features while broadening reception field. The model proposed is suitable for FER of low resolution inputs, and it can improve the effect of FER effectively. We name it Low Resolution Expression Recognition Network (LRERNet).

In addition, due to the increasing trend of the aged in Asian countries such as China and Japan, we try to monitor the expression status of the aged in the nursing homes in real time, so as to timely feedback information to the corresponding caregivers or family members, and pay attention to the health of the aged. Combining with this practical application scenario, this paper established a dataset for the study of FER of the aged, which is also used to test LRERNet. In this paper, there are two main contributions:

- A low resolution expression recognition network which is more suitable for low resolution image facial expression recognition is proposed. We test our method on three standard datasets and experiments indicate that our model achieved good results in these datasets especially on RAF [12] dataset.
- In combination with the actual application scenarios, we establish a dataset for studying FER of the old and we test our method on it.

2. Proposed Framework

We first briefly introduce dilated convolution, and then analyse the shortcomings of ordinary convolution neural networks and propose LRERNet. Finally, the loss function we used is described.

2.1. A Brief Review of Dilated Convolution

Dilated convolution [13] was originally applied to semantic segmentation tasks, and the model can provide a broader field of perception with no pooling operations and a comparable amount of computation, so as to improve the accuracy of pixel prediction. Dilated convolution can be defined as:

$$y(m,n) = \sum_{i=1}^{M} \sum_{j=1}^{N} x(m + r \times i, n + r \times j) w(i,j) \tag{1}$$

where $x(m,n)$ stands for the input information. $w(i,j)$ is a filter with a length of M and a width of N. $y(m,n)$ is the output information. The parameter r represents the dilate rate.

2.2. Low Resolution Expression Recognition Network (LRERNet)

As we all know, the performance of CNN increases with depth [14]. At the same time, for increasing the receptive field, the pool layers are usually used to subsample feature maps. However, pooling operations will lead to the loss of spatial details, so when the resolution of the images inputted directly into most CNN models is small, the depth of these CNN models is limited to some extent.

For reducing pooling operation frequency and maintain the receptive field, we improved vgg-16 by using the cavity convolution group, and designed a low resolution expression recognition network with front and back end structure. The model's structure we proposed is shown in figure 1. Referring to the related work of using VGG structure for improvement [15-18], we choose vgg-16 after removing the full connection layer as the front-end of LRERNet. The back-end of the model is composed of three dilated convolutions in series. The dilated convolution group uses dilated convolution layers with dilated rates of 1, 2 and 3 to extract the features of inputted images, and we connect these features in turn, then use 1 × 1 convolution to integrate the features of each dilated convolution to improve the expression ability. Finally, a full connection layer is connected to complete the task of classification and recognition through the Softmax layer.

2.3. Loss Function

Centre loss [19] is an improved form of Softmax loss. While monitoring the model training, the centre loss calculates a class centre for each category, and studies and corrects the class centre in the way of measuring learning, so as to increase the discriminability of features. The definition formula of the central loss function is described as

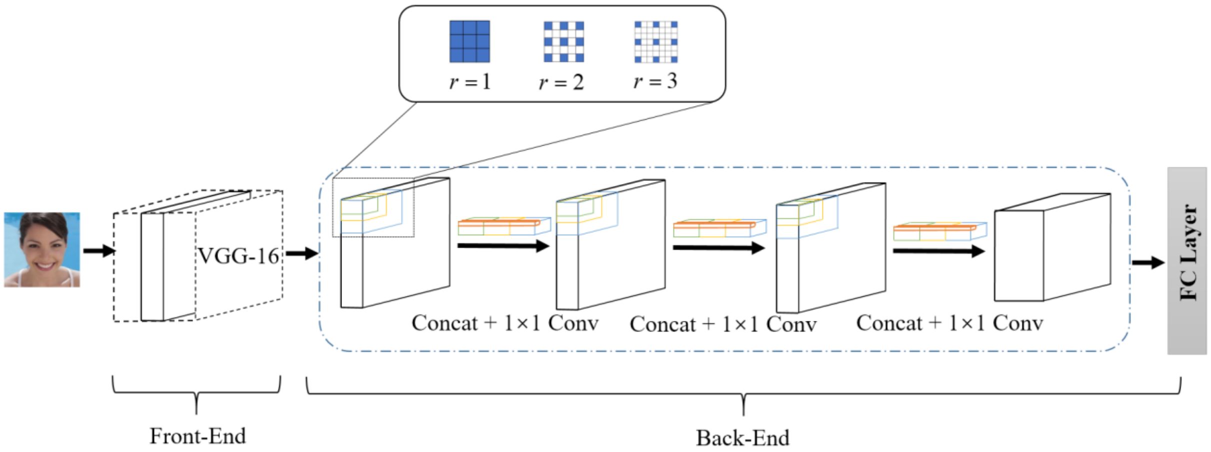


Figure 1. The structure of LRERNet.

$$L_c = \frac{1}{2}\sum_{i=1}^{n}\left\|x_i - c_{yi}\right\|_2^2 \tag{2}$$

$$\frac{\partial L_c}{\partial x_i} = x_i - c_{yi} \tag{3}$$

where n is the sample size in one batch, yi denotes the class label of the i^{th} sample, x_i is the feature vector output of the i^{th} sample, and c_{yi} is the feature centre of the category of yi.

The feature center c_{yi} is constantly updated, and the samples with similar features gradually gather to the center. The process can be described as

$$\Delta c_j = \frac{\sum_{i=1}^{n}\delta(y_i,j)\cdot(c_j - x_i)}{1+\sum_{i=1}^{n}\delta(y_i,j)} \tag{4}$$

$$c_j^{t+1} = c_j^t - \alpha\Delta c_j^t \tag{5}$$

where c_j is the feature centre of the category of j. α denotes the learning rate.

3. Experiments

We evaluated our model on several existing datasets, including RAF, SFEW [20, 21] and CK+ [22]. In addition, we build a dataset named AOPFE to study the facial expression of the elderly in Asia and evaluated our model on this dataset.

3.1. Implementation Details

First of all, we use IntraFace [23] to align the faces of the images, and save the images containing only the faces. In order to minimize the effects of various kinds of light, we transform all the inputted images into grey images. For avoiding over fitting, we expand the data using the method of horizontal flip and rotation.

The Adam optimizer is used to update parameters. We use 0.001 as the initial learning rate. And the momentum is 0.8. We implement the method proposed by Pytorch, and the GPU we use is GTX 1080TI.

3.2. Expression Recognition Results

The **RAF dataset** contains 29672 facial images in real scenes downloaded from the Internet. There are differences in illumination, head posture and face occlusion between images. At present, the image quality of RAF dataset is relatively better, the image annotation credibility is higher, and the data processing is complete. In table 1, the comparison results between LRERNet and existing methods on RAF dataset are given. From table 1 we can know that the accuracy of the LRERNet proposed in this paper is 73.62%. Compared with the baseline network and DLPCNN, the recognition effect is improved. As far as we know, the LRERNet is superior to the SOTA method on RAF dataset.

Table 1. Comparison with existing model on the RAF dataset for FER.

Model	Accuracy
VGG-16	62.97%
BaseDCNN	63.61%
DLPCNN	70.98%
LRERNet (Ours)	73.62%

The **SFEW dataset** contains different head posture and different age expression images, which can be regarded as facial expression dataset in natural environment. However, because the size of SFEW dataset is small, we use the RAF dataset to pre-train the model and do transfer training on SFEW dataset. The comparison results are shown in table 2. On the SFEW dataset, the recognition accuracy of LRERNet is 48.09%, and the recognition accuracy of the existing methods is generally low. The main reason is that there are not enough images in the dataset, and the facial expressions are all natural states, so there are many interfering factors and it is difficult to recognize. So the model needs to be further improved and improved.

Table 2. Comparison with existing model on the SFEW dataset for FER.

Model	Accuracy
AUDN [24]	30.14%
CNN-MBP [25]	51.75%
DLP-CNN	51.05%
SFEW best [26]	52.50%
GDFER [27]	47.70%
LRERNet (Ours)	48.09%

The **CK+ dataset** is widely used for FER and collected under laboratory conditions. All images in the dataset are 640×480 resolution, and the dataset contains 593 expression sequences, but only 327 of them have expression labels. For the sake of verifying the model proposed, we use part of images in the expression sequence for FER. We select peak expression images in each expression sequence as training data. From table 3, we can know that LRERNet has achieved 98.51% recognition accuracy. Compared with the RAF dataset and SFEW dataset, the overall recognition accuracy is higher. The reason is that the CK+ dataset is obtained under laboratory conditions. The face images are clear, the illumination is uniform, and the postures are unified. Therefore, the recognition rates of all methods are relatively high.

We created a facial expression dataset for the elderly in Asia named AOPFE. The images in the dataset are all facial expressions of the elderly in the natural state, and all the images are obtained through the Internet. After strict manual filtering and annotation, the final dataset contains 300 images. Most of the dataset images are happy or neutral expressions, and there are few negative expressions, so the dataset is used in the form of two classify.

We compare VGG-16 as baseline with LRERNet on this dataset. According to the results in table 4, LRERNet has achieved 86% recognition accuracy, which is significantly improved compared with the baseline model.

Table 3. Comparison with existing model on the CK+ dataset for FER.

Model	Accuracy
AUDN	93.70%
LOMo	92.00%
DLP-CNN	70.98%
DTGAN [28]	97.25%
Peak-Piloted [29]	99.30%
GDFER	93.20%
LRERNet (Ours)	98.51%

Table 4. Comparison with existing model on the AOPFE dataset for FER.

Model	Accuracy
VGG-16	84.00%
LRERNet (Ours)	86.00%

4. Conclusion

In order to reduce the number of pooling operations and maintain the receptive field, we proposed a low resolution expression recognition network, which is appropriate for low-resolution expression images. Experimental results show that the model we proposed can effectively improve the recognition ability of low-resolution expression images. Finally, we build a dataset to study the FER of the elderly and test our model on it.

References

[1] Hamm J, Kohler C G, Gur R C, et al. 2011 Automated facial action coding system for dynamic analysis of facial expressions in neuropsychiatric disorders *Journal of Neuroscience Methods* **200** (2) 237-256.

[2] Ojala T, Pietikainen M and Maenpaa T 2002 Multiresolution gray-scale and rotation invariant texture classification with local binary patterns *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** (7) 971-987.

[3] Karan S, Tingfan W, Josh S and Marian B 2012 Exploring bag of words architectures in the facial expression domain *International Conf. on Computer Vision* pp 250-259.

[4] Dalal N and Triggs B 2005 Histograms of oriented gradients for human detection *IEEE Computer Society Conf. on Computer Vision and Pattern Recognition* pp 886-893.

[5] David G L 1999 Object recognition from local scale-invariant features *IEEE International Conf. on Computer Vision* pp 1150-57.

[6] HungHsu T, YenShou L and YiCheng Z 2010 Using SVM to design facial expression recognition for shape and texture features *International Conf. on Machine Learning and Cybernetics* pp 1150-57.

[7] Zhiding Y 2015 Image based static facial expression recognition with multiple deep network learning *International Conf. on Multimodal Interaction* pp 435-442.

[8] Uddin M Z, Hassan M M, Almogren A, et al. 2017 A facial expression recognition system using robust face features from depth videos and deep learning *Computers and Electrical Engineering* **63** (1) 114-125.

[9] Kuo C, Lai S and Sarkis M 2018 A compact deep learning model for robust facial expression recognition *IEEE Conf. on Computer Vision and Pattern Recognition Workshops* pp 2202-28.

[10] Refat J and Azlan N Z 2019 Deep learning methods for facial expression recognition *International Conf. on Mechatronics Engineering* pp 1-6.

[11] Jun C, Quan C, et al. 2018 Facial expression recognition method based on sparse batch normalization CNN *37th Chinese Control Conf.* pp 9608-13.

[12] Li S, Deng W and Du J P 2017 Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild *IEEE Conf. on Computer Vision and Pattern Recognition* pp 2584-93.

[13] Yu F and Koltun V 2016 Multi-scale context aggregation by dilated convolutions *arXiv 1511.07122*.

[14] Simonyan K and Zisserman A 2014 Very deep convolutional networks for large-scale image recognition *arXiv 1409.1556*.

[15] Li Y, Zhang X and Chen D 2018 CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes *IEEE Conf. on Computer Vision and Pattern Recognition* pp 1091-1100.

[16] Boominathan L, Kruthiventi S and Babu R V 2016 CrowdNet: A deep convolutional network for dense crowd counting *ACM International Conf. on Multimedia* pp 640-644.

[17] Deepak B S, Shiv S and Babu R V 2017 Switching convolutional neural network for crowd counting *IEEE Conf. on Computer Vision and Pattern Recognition* pp 4031-39.

[18] Sindagi V A and Patel V M 2017 Generating high-quality crowd density maps using contextual pyramid CNNs *IEEE International Conf. on Computer Vision* pp 1879-88.

[19] Yandong W, Kaipeng Z, Zhifeng L and Yu Q 2016 A Discriminative Feature Learning Approach for Deep Face Recognition *European Conf. on Computer Vision* pp 499-515.

[20] Dhall A, Murthy O R, Goecke R, et al. 2015 Video and image based emotion recognition challenges in the wild: EmotiW *ACM International Conf. on Multimodal Interaction* pp 423-426.

[21] Dhall A, Goecke R, Lucey S and Gedeon T 2011 Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark *IEEE International Conf. on Computer Vision Workshops* pp 2106-12.

[22] Valstar M F, Mehu M, Jiang B, et al. 2012 Meta-analysis of the first facial expression recognition challenge *IEEE Transactions on Systems* **42** (4) 966-979.

[23] Xiong X and Torre F D 2013 Supervised descent method and its applications to face alignment *IEEE Conf. on Computer Vision and Pattern Recognition* pp 532-539.

[24] Liu M, Li S, Shan S, et al. 2015 AU-inspired deep networks for facial expression feature learning *Neurocomputing* **159** (1) 126-136.

[25] Levi G and Hassner T 2015 Emotion recognition in the wild via convolutional neural networks and mapped binary patterns *International Conf. on Multimodal Interfaces* pp 503-510.

[26] Kim B K, Roh J, Dong S Y, et al. 2016 Hierarchical committee of deep convolutional neural networks for robust facial expression recognition *Journal on Multimodal User Interfaces* **10** (2) 173-189.

[27] Mollahosseini A, Chan D and Mahoor M H 2016 Going deeper in facial expression recognition using deep neural networks *IEEE Winter Conf. on Applications of Computer Vision* pp 1-10.

[28] Jung H, Lee S, Yim J, et al. 2015 Joint fine-tuning in deep neural networks for facial expression recognition *IEEE International Conf. on Computer Vision* pp 2983-91.

[29] Zhao X, Liang X, Liu L, et al. 2016 Peak-piloted deep network for facial expression recognition *European Conf. on Computer Vision* pp 425-442.