Cognitive Neurodynamics (2022) 16:847–858
https://doi.org/10.1007/s11571-021-09761-3

RESEARCH ARTICLE

Hierarchical scale convolutional neural network for facial expression recognition

Xinqi Fan[1] · Mingjie Jiang[1] · Ali Raza Shahid[1,2] · Hong Yan[1]

Received: 26 March 2021 / Revised: 7 September 2021 / Accepted: 22 November 2021 / Published online: 5 January 2022

Abstract

Recognition of facial expressions plays an important role in understanding human behavior, classroom assessment, customer feedback, education, business, and many other human-machine interaction applications. Some researchers have realized that using features corresponding to different scales can improve the recognition accuracy, but there is a lack of a systematic study to utilize the scale information. In this work, we proposed a hierarchical scale convolutional neural network (HSNet) for facial expression recognition, which can systematically enhance the information extracted from the kernel, network, and knowledge scale. First, inspired by that the facial expression can be defined by different size facial action units and the power of sparsity, we proposed dilation Inception blocks to enhance kernel scale information extraction. Second, to supervise relatively shallow layers for learning more discriminated features from different size feature maps, we proposed a feature guided auxiliary learning approach to utilize high-level semantic features to guide the shallow layers learning. Last, since human cognitive ability can progressively be improved by learned knowledge, we mimicked such ability by knowledge transfer learning from related tasks. Extensive experiments on lab-controlled, synthesized, and in-the-wild databases showed that the proposed method substantially boosts performance, and achieved state-of-the-art accuracy on most databases. Ablation studies proved the effectiveness of modules in the proposed method.

Keywords Facial expression recognition · Hierarchical scale network · Dilated inception blocks · Feature guided auxiliary learning · Knowledge transfer learning

Introduction

Facial expression recognition (FER) refers to infer human expressions or emotions from their face images (Li and Deng 2020). Emotions can also be identified by electroencephalogram (EEG) signals (Chen et al. 2015; Shen et al. 2020). Many real-world applications are based on FER. FER has been used to collect implicit feedback from customers (Kasiran and Yahya 2007), which could help administrators understand reviews and improve their operating strategies. In education, a teaching feedback collection system based on FER has been developed to visualize students' emotions in classrooms (Zeng et al. 2020). Many other applications for medical treatment (Shih et al. 2008), surveillance (Ocegueda et al. 2011), autonomous driving and other human-computer interaction applications have been developed (Bartlett et al. 2003). However, the study of FER, which is currently under active research, still needs improvement. The most widely used quantitative measurement of FER is categorical facial expressions (Mollahosseini et al. 2017). Categorical facial expressions usually contain seven emotions with six basic emotions, angry, disgusted, fearful/afraid, happy, sad, surprised, and one neutral emotion (Wang et al. 2010). In Li et al. (2017), the contemptuous is also included as an emotion category.

A general pipeline for FER can be summarized as detecting faces, aligning faces, extracting facial expression features, and classifying expressions (Li and Deng 2020). In most research studies on FER, face detection is performed separately, so researchers focus on extracting

✉ Xinqi Fan
xinqi.fan@my.cityu.edu.hk

[1] Department of Electrical Engineering, City University of Hong Kong, Hong Kong SAR, China

[2] Electrical and Computer Engineering Department, COMSATS University Islamabad, Islamabad, Pakistan

features related to facial expression and recognition of expressions. Many researchers designed hand-crafted feature extractors followed by a support vector machine (SVM) or decision tree (DT) as a classifier to obtain the final expressions (Bartlett et al. 2005). Deep learning is changing the development of FER, leading to methodologies with higher performance (Li et al. 2018; Wen et al. 2020; Bai et al. 2009). In Mollahosseini et al. (2016), the researchers introduced a convolutional neural network (CNN) containing four Inception modules to extract FER related features using 1×1, 3×3 and 5×5 kernels and claimed that the sparsity of the Inception helped to learn, but they still used the original Inception blocks, which increased the model size. Hu et al. (2017) applied deep supervision to intermediate feature maps for FER, but high-level semantic information was not used for intermediate layers learning.

In this study, we proposed a hierarchical scale convolutional neural network (HSNet) for FER to systematically enhance information extracted from the kernel, network, and knowledge scales. First, inspired by that the facial expression that can be defined by different size facial action units (AUs) and the power of sparsity, we proposed dilation Inception blocks (DIBs), which contains a standard dilation Inception block (SDIB) and a residual dilated Inception block (RDIB), to enhance kernel scale information extraction. Second, to supervise relatively shallow layers for learning more discriminated features from different size feature maps, we proposed a feature guided auxiliary learning (FGAL) to utilize high-level semantic information to guide the shallow layers' learning. Last, since human cognitive ability can progressively be improved by learned knowledge, we mimicked such ability by knowledge transfer learning (KTL) to use knowledge learned from related tasks. Extensive experiments on lab-controlled, synthesized, and in-the-wild databases showed that the proposed method substantially boosts performance, and achieved state-of-the-art accuracy on most databases. Ablation studies proved the effectiveness of modules in the proposed method. The proposed HSNet obeys a general hierarchy, which is better visualized by a scale pyramid in Fig. 1. Sample images from the used databases are shown in Fig. 2 with details given in Sect. 4.1. In summary, the main contributions of this work include:

- We designed an HSNet with a hierarchical scale idea for enhancing information extraction from the kernel, network, and knowledge scales.
- We proposed DIBs with different receptive field kernels to extract features with a sparser architecture than the original Inception modules.
- We utilized high-level semantic information to guide the learning of intermediate layers through FGAL.
- Superior performance was shown on lab-controlled, synthesized, and in-the-wild databases. HSNet achieved the state-of-the-art accuracy on RaFD (99.88%).

The rest of this paper is organized as follows. In Sect. 2, the general background of facial expression recognition, and related deep learning basics are presented. The proposed HSNet is presented in Sect. 3. Section 4 describes the experiments and their results. Section 5 concludes the paper with potential future work.

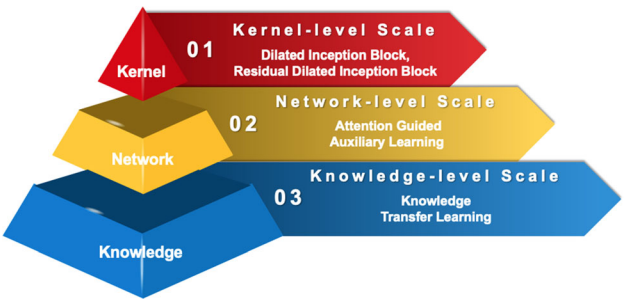


Fig. 1 Scale pyramid for extracting information from the kernel, network, and knowledge scales for deep neural networks

Related work

Facial expression recognition

Traditional machine learning methods for FER focus on face geometry, edges, and contours to extract facial features. Shan et al. (2005) proposed a low computation FER algorithm based on the local binary pattern (LBP) algorithm, but it did not achieve satisfying performance when dealing with faces in different orientations. Berretti et al.

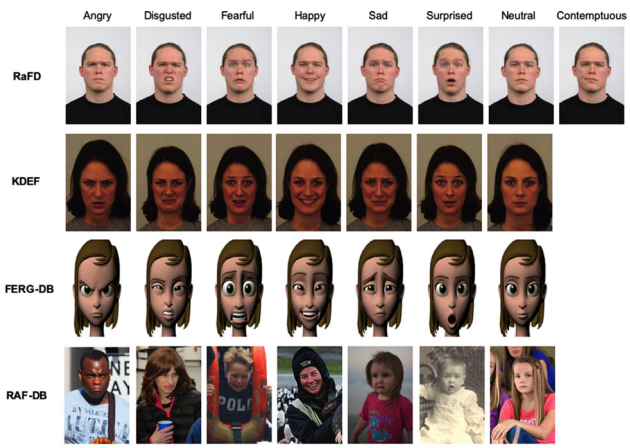


Fig. 2 Sample images from facial expression databases. RaFD database has eight expressions, while KDEF, FERG-DB, and RAF-DB have seven expressions each. RAF-DB has challenging images, in which some expressions are even harder for humans to recognize

(2010) extracted and selected scale invariant feature transform (SIFT) features for FER to improve extracted features in different scales, but it had a relatively high computational cost. Contour and region harmonic features for sub-local FER were proposed by Shahid et al. (2020). Avani et al. (2020) pointed out lips geometry features are useful for facial expression recognition, so they proposed a method using lips features based on the properties of parabola. To select the most relevant handcrafted features, singular vector decomposition (SVD) based co-clustering in feature extraction was adopted for FER to reveal the salient features in terms of attention maps and enhanced the learning performance (Khan et al. 2016, 2017).

Deep learning based FER has become popular over the past few years, because it can extract features of good quality and perform tasks in an end-to-end manner (Li and Deng 2020). CNNs are used in most image based FER applications to extract features, while recurrent neural networks (RNNs) are used in sequence and video based FER studies (Li and Deng 2020). Khorrami et al. (2015) indicated that learned CNN features have a high relation with facial AUs. Barsoum et al. (2016) pointed out that there are noisy labels in the FER databases, but did not provide an efficient solution. Balahur et al. (2011) introduced a deep learning based network called EmotiNet to recognize emotion with text built on appraisal theories, but it was limited by concept nuances and complex linguistic rules. In Fan et al. (2020), a hybrid separable convolutional inception residual network (HSCIRN) for FER was developed, which reduces the number of trainable parameters for embedded system friendly applications, but the method lost accuracy. Due to the larger intra-class variances than inter-class variances in some cases, center loss based-CNN (Center-CNN) by Wen et al. (2016) and deep locality-preserving CNN (DLP-CNN) by Li et al. (2017) explicitly minimized the intra-class variances to overcome such issue. Zhang et al. (2020) claimed that identity information could help the FER and developed an identity-expression dual branch network (IE-DBN) to improve the recognition performance. However, none of these methods considered to extract robust features systematically for FER.

Convolutional neural network

The winner of the ImageNet challenge, AlexNet, demonstrated the power of CNN, when there are enough computational resources and data (Krizhevsky et al. 2012). However, these networks achieve limited feature extraction due to their limited number of layers. A deeper network was achieved by Inception, in which the network learns to select different kernels, 1×1, 3×3, 5×5 and pooling, rather than have them selected by designers (Szegedy et al. 2015). Different kernels represent various receptive field sizes, so Inception is capable of multiple scale feature extraction. Inception networks were improved by replacing one 5×5 kernel as two 3×3 kernels in Inception-V2 and replacing $n \times n$ by its asymmetric pairs $1 \times n$ and $n \times 1$ kernels in Inception-V3 (Szegedy et al. 2016). Residual network (ResNet) utilized a skip connection to solve the degradation problem, and allowed the network to be much deeper (He et al. 2016). In addition, to improve the effectiveness and making Inception deeper, Inception-v4 and Inception-ResNets were also developed by Szegedy et al. (2017).

The dilated convolution (a.k.a atrous convolution) operations insert holes among convolution kernels, so that they can enlarge the receptive field and extract better contextual information without increasing the number of parameters (Yu and Koltun 2016). Compared with standard convolutions

$$F^l(i,j) = \sum_m \sum_n F^{l-1}(i+m, j+m)\mathcal{K}(m,n), \tag{1}$$

dilated convolution has skips in its operation as

$$\begin{aligned} F^l(i,j) &= (F^{l-1} \otimes \mathcal{K}_d)(i,j) \\ &= \sum_m \sum_n F^{l-1}(i+m\times d, j+n\times d)\mathcal{K}(m,n), \end{aligned} \tag{2}$$

where F^l and F^{l-1} are feature maps at layer l and $l-1$ where $l \in \{1, 2, \ldots, L\}$, $\mathcal{K}$ is the standard convolution filter, $\mathcal{K}_d$ is the dilated convolution filter and d is the dilation rate. It can be seen that the dilated convolution does not increase the number of learnable parameters in kernels. For a standard convolution with size $k \times k$, the receptive field of the dilated convolution kernel with dilation d becomes $k_d \times k_d$, where

$$k_d = k + (d-1)(k-1). \tag{3}$$

Therefore, by increasing the dilation rate d, the receptive field of a kernel can be changed, while the number of parameters remains unchanged as the ordinary kernels.

Knowledge transfer learning

Deep learning requires a vast amount of data to learn latent patterns, but in many cases, collecting and annotating data is difficult (Tan et al. 2018). KTL, which can transfer the knowledge from the source domain to the target domain, becomes very useful in these cases for reducing the effort to collect and annotate more data. Given a source domain and source task, and a target domain and target task, KTL intends to assist in the learning of the target function utilizing knowledge from the source domain and source

task (Pan and Yang 2009). There are two categories of KTL, inductive and transductive one (Deng et al. 2014). The source and target tasks are the same in the former setting, while they are different in the latter. In deep learning based KTL, most methods focused on the inductive KTL. Neural networks usually learn common features in the first few layers, while they learn specific features in the last layers (Yosinski et al. 2014). In addition, knowledge learned from tasks with a closer relationship can improve the model performance (Zamir et al. 2018). Recently, there are several successful applications of KTL (Chen et al. 2020; Devlin et al. 2019; Abbasi et al. 2020). The KTL process is similar to the human cognitive ability that can progressively be improved by the knowledge learned from other tasks. Inspired by the similarity between KTL and how humans improve cognitive ability, we employed KTL to transfer useful knowledge for the FER task.

Methodology

Given a 3D tensor $\mathbf{x}$ as an input facial image, and the ground truth expression scalar label $\mathbf{y} \in \{0, \ldots, C-1\}$, we aim to use the proposed HSNet to predict a probabilistic expression vector $\hat{Y}$ where the index of the highest probability acts as the predicted class. To achieve this, HSNet utilizes a hierarchical way to enhance the feature extraction ability from the kernel (Sect. 3.1), network (Sect. 3.2), and knowledge scales (Sect. 3.3). Finally, we summarize the network structure and present the loss function in Sect. 3.4. The schematic of HSNet is shown in Fig. 3.

Kernel-scale enhancement: dilated inception blocks

The motivation for using dilated Inception comes from facial expressions being represented by a combination of facial AUs in different sizes (Lucey et al. 2010). Therefore, novel DIBs (Fig. 4), consisted of SDIBs and RDIBs, were proposed to extract features using kernels with different receptive fields. Compared with the original Inception module, DIBs not only have kernels with different receptive fields, but also have reduced the number of parameters of the network, which can result in a sparser representation.

The general architecture of SDIBs is shown in Fig. 4a, where the block uses 3×3 convolutions with different dilation rates d to realize the different receptive fields, 1×1 convolutions to decrease the number of channels, and a max pooling to preserve the orientation-invariant property. By assuming the spatial input size as $H \times W$, the spatial output size is half of the spatial input size subtracted by one rounded down to an integer as $\lfloor (H-1)/2 \rfloor \times \lfloor (W-1)/2 \rfloor$, while the number of channels depends on the specific implementation. Let S_i^l be the intermediate feature map for ith branch of the layer l. SDIB can be described as

$$S_i^l = F^{l-1} \otimes \mathcal{K}_{1\times1}^l \otimes \mathcal{K}_{3\times3,d}^l \tag{4}$$

$$(i \in \{1,2,3\}, d \in \{\alpha, \beta, \gamma\}),$$
$$S_4^l = F^{l-1} \otimes \mathcal{P}_{3\times3}^l \otimes \mathcal{K}_{1\times1}^l, \tag{5}$$

$$F^l = S_1^l \oplus S_2^l \oplus S_3^l \oplus S_4^l, \tag{6}$$

where $\mathcal{K}_{k\times k}$ and $\mathcal{P}_{k\times k}$ refer to the $k \times k$ convolution and max-pooling kernels; α, β and γ are the dilation rates; $\otimes$

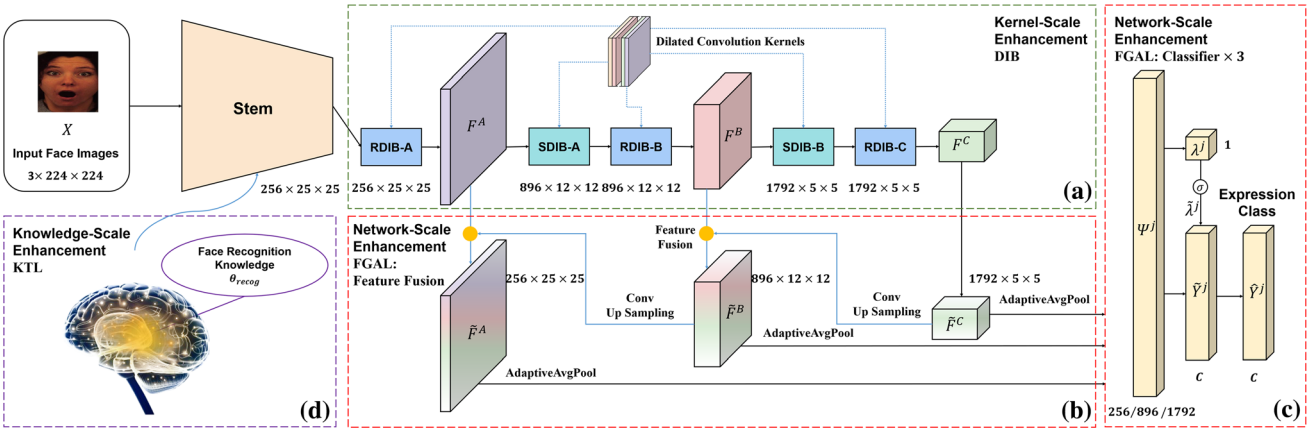
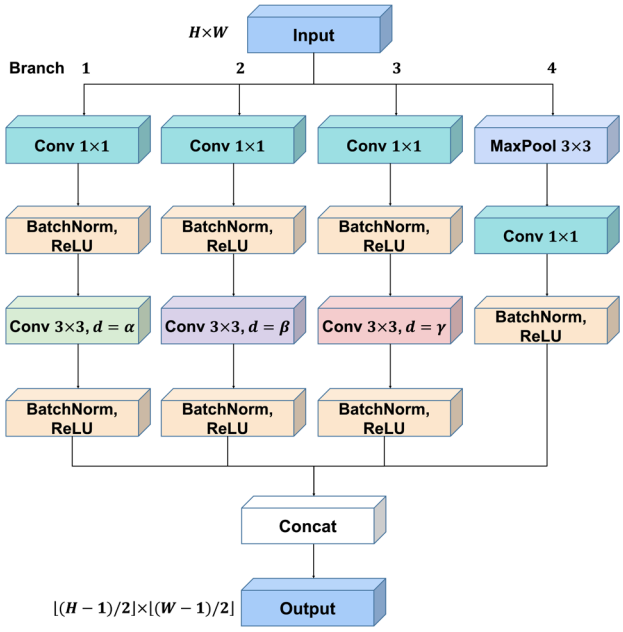
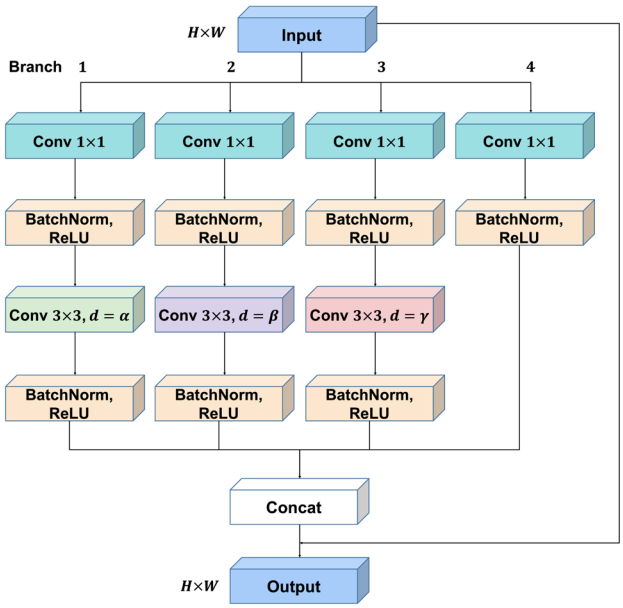


Fig. 3 Schematic of HSNet. **a** shows kernel-scale enhancement which is achieved by DIBs consisted of SDIBs and RDIBs in the dashed green box. **b** shows the first part of network-scale enhancement which is realized by the feature fusion of FGAL using high-level semantic features in the dashed red box. **c** shows the second part of network-scale enhancement, which is implemented by self-guided classifiers of FGAL to learn weights for all classifiers in the dashed red box. **d** shows knowledge-scale enhancement which is performed by KTL from face recognition in the dashed purple box. The input size is assumed as $3 \times 224 \times 224$, but it can be varied



(a) Architecture of SDIBs



(b) Architecture of RDIBs

Fig. 4 General architectures of DIBs. Conv stands for the convolution, and MaxPool for the max pooling

and $\oplus$ denote the convolution and concatenation operations, respectively; Batch normalization and rectified linear units (ReLUs) are ignored in the expressions for simplicity.

Similarly, RDIBs (Fig. 4b) are realized by 3×3 convolutions with different dilation rates d, and 1×1 convolutions. In addition to that, a skip connection has been used to add the input feature map F^{l-1} and the concatenated feature map S^l to obtain the final output feature map as

$$S_i^l = F^{l-1} \otimes \mathcal{K}_{1\times1}^l \otimes \mathcal{K}_{3\times3,d}^l \quad (i \in \{1,2,3\}, d \in \{\alpha,\beta,\gamma\}), \tag{7,8}$$

$$S_4^l = F^{l-1} \otimes \mathcal{K}_{1\times1}^l,$$

$$S^l = S_1^l \oplus S_2^l \oplus S_3^l \oplus S_4^l, \tag{9}$$

$$F^l = F^{l-1} + S^l, \tag{10}$$

where the output feature map size is the same as the input feature map size.

During the implementation, there are variations for each SDIB and RDIB as shown in Fig. 3a via setting different dilation rates, and excluding some branches.

Network-scale enhancement: feature guided auxiliary learning

The auxiliary learning for intermediate layers has been used in Szegedy et al. (2015) to avoid gradient vanishing problem, but the high-level semantic information is not used. We proposed a novel FGAL approach to let the network learn more discriminate features from intermediate feature maps by utilizing high-level semantic feature maps. This will enable the network to learn more fine-grained features from bigger feature maps.

Inspired by feature pyramid network (FPN) from the object detection task (Lin et al. 2017) to incorporate high-level semantic features, FGAL uses deeper features as high-level features and generates fused feature maps for different scales (Fig. 3b). After each RDIB, the network obtains a low-level feature map F^A, a medium-level feature map F^B and a high-level feature map F^C. First, an 1×1 convolution is applied on F^C to obtain $\tilde{F}^C$ as

$$\tilde{F}^C = F^C \otimes \mathcal{K}_{1\times1}^C. \tag{11}$$

Then, the high-level feature map $\tilde{F}^C$ is convolved by a convolution to adjust the number of channels, and a nearest upsampling operation $\mathcal{UP}$ to enlarge the size. Finally, it is added with the feature map F^B to obtain $\tilde{F}^B$ as

$$\tilde{F}^B = F^B + \mathcal{UP}(\tilde{F}^C \otimes \mathcal{K}_{1\times1}^B), \tag{12}$$

Similarly, we can obtain $\tilde{F}^A$ from F^A and $\tilde{F}^B$ as

$$\tilde{F}^A = F^A + \mathcal{UP}(\tilde{F}^B \otimes \mathcal{K}_{1\times1}^A). \tag{13}$$

Since there are three fused features maps, but one classifier is needed for the model's final output, FGAL learns a weight for each level feature map to guide itself for indicating its importance (Fig. 3c). This self-guided classifier using its own fused feature map is inspired by attention in natural language processing (Luong et al. 2015). Consider

one of the fused feature maps as $\tilde{F}^j$ ($j \in \{A, B, C\}$), it first goes through an adaptive average pooling and then resizes as a feature vector Ψ^j, where the size depends on the number of channels of $\tilde{F}^j$. Then, there are two fully connected (FC) layers, one $\mathcal{W}_1^j$ for mapping to a scalar followed by a sigmoid function σ to obtain a learned weight $\tilde{\lambda}_j$, and one $\mathcal{W}_C^j$ for mapping to an unweighted class output $\tilde{Y}^j \in C$. The weighted class output $\hat{Y}^j$ for the jth classifier is obtained by multiplying the weight with the unweighted class output as

$$\hat{Y}^j = \sigma(\Psi^j \otimes \mathcal{W}_1^j) \cdot (\Psi^j \otimes \mathcal{W}_C^j) \tag{14}$$

$$= \sigma(\lambda_j) \cdot \tilde{Y}^j = \tilde{\lambda}_j \cdot \tilde{Y}^j. \tag{15}$$

Finally, all the three classifiers' outputs $\hat{Y}^A$, $\hat{Y}^B$, $\hat{Y}^C$ are summed to obtain the model's final output as

$$\hat{Y} = \tilde{\lambda}_A \cdot \tilde{Y}^A + \tilde{\lambda}_B \cdot \tilde{Y}^B + \tilde{\lambda}_C \cdot \tilde{Y}^C \tag{16}$$

$$= \hat{Y}^A + \hat{Y}^C + \hat{Y}^C. \tag{17}$$

Knowledge-scale enhancement: knowledge transfer learning

Inspired by how humans improve cognitive ability, we applied KTL (Fig. 3d) to enlarge the knowledge scale. First, an Inception-ResNetV1 (Szegedy et al. 2017) model was used as the source model, which was trained on the source face recognition domain using the VGGFace2 dataset with 3.31 million images to learn general face related knowledge (Cao et al. 2018). The learned knowledge for the stem is denoted as θ_{recog}. Then, the proposed HSNet $f(\theta)$ is used as the target model for the target FER domain, consisting of the same stem as the Inception-ResNetV1, and the subsequent modules as discussed in the above sections. The parameter of HSNet θ consists of 2 parts as θ_s for the stem, and θ_h for all the rest parts. Then, the KTL starts by initializing the stem parameters from the face recognition knowledge and then randomly initializing the rest as

$$f(\theta) = f(\theta_s, \theta_h)|_{\theta_s \leftarrow \theta_{recog}; \theta_h \leftarrow HeInit(\theta_h)}, \tag{18}$$

where *HeInit* stands for the He initialization (He et al. 2015). Assuming the loss is $\mathcal{L}$, when updating the parameters, θ_s is frozen as the fixed knowledge and θ_h is updated using the gradient descent of learning rate ρ as

$$\theta_s \leftarrow \theta_{recog}, \tag{19}$$

$$\theta_h \leftarrow \theta_h - \rho \nabla_{\theta_h} \mathcal{L}. \tag{20}$$

Network structure and loss function

The complete HSNet structure is built upon each component as mentioned in the previous section except for the stem network used from the Inception-ResNetV1 (Szegedy et al. 2017). The HSNet structure details are defined in Table 1, where we include all convolutions, poolings, upsampling, and FC layers, but omit BatchNorm and activation functions for simplicity. The input image size is assumed as $3 \times 224 \times 224$ in the Table 1 and Fig. 3 for demonstration purposes, but the size can be varied.

For training, the maximum likelihood estimation (MLE) is used to derive the the multi-class cross-entropy (CE) as the loss function. Consider a minibatch of training samples $\{(\mathbf{x}_i, \mathbf{y}_i) \mid i = 1, \ldots, N_B\}$, and the model generates a predicted vector $\hat{Y}_i = f(\mathbf{x_i}; \theta)$. The loss can be expressed as

$$\mathcal{L} = -\frac{1}{N_B} \sum_{i=1}^{N_B} \log \left(\frac{e^{\hat{Y}_i(\mathbf{y}_i)}}{\sum_{j=1}^{C} e^{\hat{Y}_i(j)}} \right), \tag{21}$$

where N_B is the number of samples within a minibatch.

Experiment and result

Comprehensive experiments were conducted on frequently used databases. We also performed ablation studies for the proposed modules.

Database

RaFD

The Radboud Faces Database (RaFD) (Langner et al. 2010) is a laboratory-controlled FER database covering 8 expressions as 6 basic, 1 neutral and 1 contemptuous expressions. It has 8056 images of 67 subjects.

KDEF

The Karolinska Directed Emotional Faces (KDEF) (Lundqvist et al. 1998) is an early laboratory-controlled FER database with 7 expressions as 6 basic and 1 neutral expressions. It contains 4,900 images taken from 70 subjects who were well informed and instructed before image acquisition, and 5 images were taken from different views.

FERG-DB

The Facial Expression Research Group 2D Database (FERG-DB) (Aneja et al. 2016) is a synthesized character style database containing 7 expressions, 6 basic and 1

Table 1 The details of the network structures

Module	Operation				Output Size
Input	-				$3 \times 224 \times 224$
Stem	Conv 3×3		MaxPool 3×3		
	2, 0, 1		-		$32 \times 111 \times 111$
	1, 0, 1		-		$32 \times 111 \times 111$
	1, 1, 1		-		$32 \times 109 \times 109$
	1, 0, 1		-		$34 \times 109 \times 109$
	-		2, 0, 1		$64 \times 54 \times 54$
	1, 0, 1		-		$80 \times 54 \times 54$
	1, 0, 1		-		$192 \times 52 \times 52$
	2, 0, 1		-		$256 \times 25 \times 25$
RDIB-A	Branch	Conv 3×3	Conv 1×1	MaxPool 3×3	
	1	1, 1, 1	1, 0, 1	-	$256 \times 25 \times 25$
	2	2, 2, 1	1, 0, 1	-	
	-	-	-	-	
	4	-	-	1, 0, 1	
SDIB-A	1	2, 0, 1	1, 0, 1	-	$896 \times 12 \times 12$
	2	2, 1, 2	1, 0, 1	-	
	-	-	-	-	
	4	-	1, 0, 1	2, 0, 1	
RDIB-B	1	1, 1, 1	1, 0, 1	-	$896 \times 12 \times 12$
	2	1, 2, 2	1, 0, 1	-	
	3	1, 3, 3	1, 0, 1	-	
	4	-	1, 0, 1	-	
SDIB-B	1	2, 0, 1	1, 0, 1	-	$1792 \times 5 \times 5$
	2	2, 1, 2	1, 0, 1	-	
	3	2, 2, 3	1, 0, 1	-	
	4	-	1, 0, 1	2, 0, 1	
RDIB-C	1	1, 1, 1	1, 0, 1	-	$1792 \times 5 \times 5$
	-	-	-	-	
	-	-	-	-	
	4	1, 0, 1	-	-	
FGAL	Branch	UpSample	Conv 1×1	AdaptAvgPool	FC	
	A	nearest	1, 0, 1	1	1; C	C
	B	nearest	1, 0, 1	1	1; C	C
	C	nearest	1, 0, 1	1	1; C	C
Output	-					C

For the convolution (Conv) and max pooling (MaxPool) operations, the number besides is the kernel size and the numbers under them are the stride, zero padding and dilation rate. For the adaptive average pooling (AdaptAvgPool) and fully-connected (FC) layer, the numbers below are the output dimensions. The nearest mode is used for upsampling (UpSample). The input image size is assumed as $3 \times 224 \times 224$

neutral expressions, based on 6 cartoon characters. The 55,767 annotated faces were created by MAYA (Trepagnier et al. 2006) using CNN models trained on the Extended Cohn-Kanade (CK+), Denver Intensity of Spontaneous Facial Actions (DISFA), MMI and KDEF databases.

RAF-DB

The Real-world Affective Face Database (RAF-DB) is an FER database collected from the "wild" (Li et al. 2017; Li and Deng 2018). The database has 7 expressions, 6 basic emotions and neutral, with 29,672 images covering diverse poses, illuminations, and backgrounds. It facilitates research from lab-controlled conditions to natural or "in-the-wild" environments.

Implementation detail

The experiments were implemented using the PyTorch deep learning framework (Paszke et al. 2019). An adaptive moment (Adam) optimizer (Kingma and Ba 2015) was used with an initial learning rate of 0.01, and the first and second order momentums of 0.9 and 0.999 respectively. The model was trained for 500 epochs with a minibatch size of 32. The ReduceLROnPlateau scheduler dynamically reduced the learning rate if the validation loss does not change for 20 epochs. Random horizontal flips and 5 degree rotations were applied as the data augmentation. Experiments were conducted on a computer with an Intel Xeon Silver 4108 CPU and an NVIDIA GeForce RTX 2080Ti GPU. For data pre-processing, we used multitask cascaded convolutional networks (MTCNN) to detect faces (Zhang et al. 2016). For RaFD and KDEF, faces were detected and cropped to 100×100 and 256×256 resolution, while other datasets provide cropped faces and face detection is not needed.

Evaluation metrics

As a classification task, it is normally evaluated by using the top-1 overall accuracy (Acc.) as

$$\text{Acc.} = \frac{\sum_{j=1}^{C} TP_j}{N}, \tag{22}$$

where TP_j stands for truth positive for class j; $N = \sum_{j=1}^{C} N_j$ is the sum of the number of samples for each class N_j. A common evaluation metric for imbalanced data is the mean accuracy (mAcc.), which takes the mean of accuracy for each class as

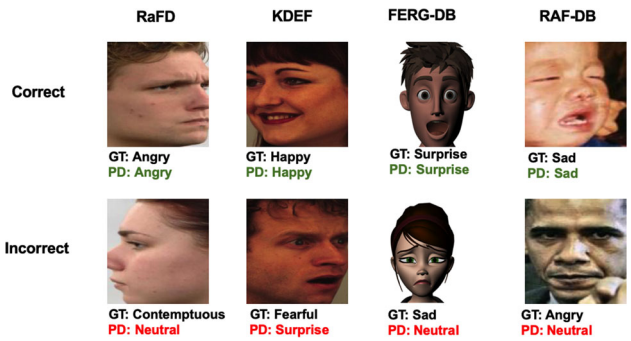


Fig. 5 Demonstration of qualitative recognition results for each database. GT: ground truth label; PD: predicted label

$$\text{mAcc.} = \left(\sum_{j=1}^{C} \frac{TP_j}{N_j}\right)/C. \tag{23}$$

Experiment and discussion

Demonstration of qualitative results is shown in Fig. 5. Quantitative results including overall accuracy, mean accuracy and confusion matrices are discussed below.

Experiment on RaFD

No face alignment was applied, since some images were taken completely from side views. The database was divided into training, validation and test sets of 5,152, 1,440, and 1,608 images respectively. Each expression class was balanced with 644 images for training, 160 for validation and 201 for testing.

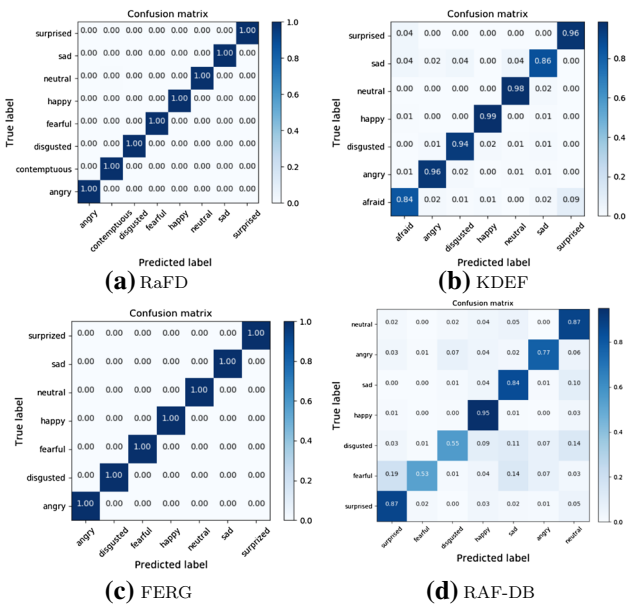


Fig. 6 Confusion matrix of test sets of different datasets

The confusion matrix of the experiments on the RaFD database is shown in Fig. 6a. Table 2 shows that HSNet achieved 99.88% predicted accuracy, which is the state-of-the art result on RaFD. HSNet, through its hierarchical scale feature extraction process, has excellently learned features to discriminate most expressions. Some errors were made on the neutral, contemptuous, fear and disgust classes, however, these expressions, especially neutral and contemptuous, are confusing for people in some cases.

Experiment on KDEF

For training 7798 images were selected and divided into training and validation sets at an 8:2 ratio. 1956 images were selected for testing. The number of images for each expression class in this database is almost equal.

The performance of HSNet and results from other advanced methods are summarized in Table 3. HSNet outperforms the other methods, achieving an overall accuracy of 93.35%. The confusion matrix for HSNet is shown in Fig. 6b and it indicates that the expressions of afraid and surprised can be confused, as approximately 10% of afraid expressions are recognized as surprised. This may due to the main characteristic of both expressions being the opened mouth.

Experiment on FERG-DB

In this synthetic database, the backgrounds of images are pure white or black, so the original images with resolution 256×256 can be used to train the network. The purpose to test on this dataset is that more and more people tend to replace their faces as characters on social media to protect their privacy, so the expression recognition of those synthesized faces becomes important.

The performance of HSNet and other advanced methods is summarized in Table 4. FERG-DB is a relatively easy database for FER, since the synthesized faces do not have a dense representation of facial expressions and most methods achieve high accuracy. HSNet still obtained the highest accuracy at 99.35%. The confusion matrix for HSNet on FERG-DB is shown in Fig. 6c.

Experiment on RAF-DB

The RAF-DB database provides aligned images and has pre-defined training and testing sets of 12,271 and 3065 images, respectively. The training data was categorized at a 9:1 ratio for training and validation. This "in-the-wild" database is highly imbalanced. Of the 12,271 original training images, 1290 are for surprised, 1982 for sad, 2524 for neutral, 4772 for happy, 281 for fearful, 705 for angry and 717 images for disgusted expressions. The input image is resized as 224×224.

Table 2 Accuracy on RaFD database

Method	Acc. (%)
Bayesian (Mao et al. 2016)	74.96
ExpNet (Chang et al. 2018)	75.00
DAECNN (Prieto and Oplatkova 2018)	78.51
VSDL (Mavani et al. 2017)	95.71
HSCIRN (Fan et al. 2020)	96.87
SBNN (Sun et al. 2017)	96.93
DeepExp3D (Koujan et al. 2020)	97.65
HSNet	99.88

The performance of HSNet and other advanced methods is summarized in Table 5. As this database is an imbalanced database, we evaluate the performance based on both overall accuracy and mean accuracy against each class. From the confusion matrix (Fig. 6d), HSNet performs well on the surprised, happy, sad, and neutral expressions, but did poorly on the angry, fearful and disgusted expressions. The decrease of accuracy on angry, fearful and disgusted expressions is likely because there are fewer images in the training database for these classes. Despite not using advanced methods to reduce intra-class and increase inter-class distances, HSNet outperformed DLP-CNN in mean accuracy by 2.69%. This may due to that our network can learn fine-grained discriminated features from multiple scales. In addition, HSNet surpassed a recent method IE-DBN by approximately 2% in overall accuracy. The result proves that HSNet are able to generalize to the in-the-wild scenarios.

Ablation study

An ablation study on the RAF-DB database was conducted for the DIB, FGAL and KTL components, which correspond to the kernel, network, and knowledge scales enhancement. The result is shown in Table 6 and described below. The baseline network is Inception-ResNet v1 presented in Szegedy et al. (2017).

First, by adding a single module, DIB, FGAL or KTL, into the baseline network, the accuracy increased by no more than 1%. The increase caused by DIB is surprising. This improvement may not only due to the different receptive fields created by the DIB, but also the sparser representation yielded from the dilation operations. Second, by adding two modules to the baseline, we can observe an increase of more than 1%. Among these results, we see that FGAL may play a slightly more important role for contributing to the final improvement. This

Table 3 Accuracy on KDEF database

Method	Acc. (%)
ExpNet (Chang et al. 2018)	71.00
FDCNN (Zavarez et al. 2017)	72.55
HOG-SRC (Ali et al. 2017)	78.00
SDAE (Ruiz-Garcia et al. 2017)	86.73
AFER (Yaddaden et al. 2018)	90.62
DeepExp3D (Koujan et al. 2020)	92.24
HSNet	93.35

Table 4 Accuracy on FERG-DB database

Method	Acc. (%)
DeepExpr (Aneja et al. 2016)	89.02
EMF (Zhao et al. 2018)	97.00
ANN (Feutry et al. 2018)	98.20
Deep-Emotion (Minaee and Abdolrashidi 2019)	99.30
HSNet	99.35

Table 5 Accuracy on RAF-DB database

Method	mAcc. (%)	Acc. (%)
ECNN (Lian et al. 2020)	–	82.69
DCNN (Li and Deng 2018)	72.42	82.86
Center-CNN (Wen et al. 2016)	73.42	83.86
DLP-CNN (Li and Deng 2020)	74.20	84.13
PAT-ResNet (Cai et al. 2018)	–	84.19
IE-DBN (Zhang et al. 2020)	–	84.75
HSNet	76.89	86.67

improvement shows that using high-level semantic feature to guide the learning of intermediate features are very useful in FER tasks. Last, using DIB, FGAL and KTL, we can see the accuracy has another approximately 0.5% increase. This proves that knowledge transfer can help the learning of other similar tasks, and the three modules can coordinate with each other to work as better as they can.

Conclusion

In this study, we proposed an HSNet for FER to systematically enhance information extracted from kernel, network, and knowledge scales. First, inspired by the facial expression can be defined by different size facial AUs and the power of sparsity, we proposed DIBs to enhance kernel scale information extraction. Second, to supervise relatively shallow layers for learning more discriminated features from different size feature maps, we proposed an FGAL to utilize high-level semantic information to guide the shallow layers' learning. Last, since human cognitive ability can progressively be improved by learned knowledge, we mimic such ability by KTL from related tasks. The experimental results demonstrate state-of-the-art performance on several databases, especially the 99.88% accuracy achieved on the RaFD database. Ongoing work is to design more robust methods to attack problems with FER "in-the-wild", such as the long-tailed and mislabelling problems.

Table 6 Ablation study on RAF-DB

DIB	FGAL	KTL	Acc.
×	×	×	84.54
✓	×	×	85.10
×	✓	×	85.73
×	×	✓	85.26
✓	✓	×	86.21
✓	×	✓	86.05
×	✓	✓	86.46
✓	✓	✓	86.67

Acknowledgements This work was supported by the Hong Kong Innovation and Technology Commission, and the City University of Hong Kong (Project 9610460).

Declarations

Conflict of interest The authors declare that they have no conflict of interest.

References

Abbasi AA, Hussain L, Awan IA, Abbasi I, Majid A, Nadeem MSA, Chaudhary QA (2020) Detecting prostate cancer using deep learning convolution neural network with transfer learning approach. Cogn Neurodyn 14(4):523–533

Ali AM, Zhuang H, Ibrahim AK (2017) An approach for facial expression classification. In J Biometrics 9(2):96–112

Aneja D, Colburn A, Faigin G, Shapiro L, Mones B (2016) Modeling stylized character expressions via deep learning. In: Asian conference on computer vision, springer, pp 136–153

Avani VS, Shaila S, Vadivel A (2020) Geometrical features of lips using the properties of parabola for recognizing facial expression. Cognitive Neurodyn. https://doi.org/10.1007/s11571-020-09638-x

Bai Y, Guo L, Jin L, Huang Q (2009) A novel feature extraction method using pyramid histogram of orientation gradients for smile recognition. In: IEEE International conference on image processing, IEEE, pp 3305–3308

Balahur A, Hermida JM, Montoyo A, Muñoz R (2011) Emotinet: A knowledge base for emotion detection in text built on the appraisal theories. In: International conference on application of natural language to information systems, Springer, pp 27–39

Barsoum E, Zhang C, Ferrer CC, Zhang Z (2016) Training deep networks for facial expression recognition with crowd-sourced label distribution. In: ACM international conference on multimodal interaction, ACM, pp 279–283

Bartlett MS, Littlewort G, Fasel I, Movellan JR (2003) Real time face detection and facial expression recognition: Development and applications to human computer interaction. In: IEEE conference on computer vision and pattern recognition workshop, IEEE 5:53–53

Bartlett MS, Littlewort G, Frank M, Lainscsek C, Fasel I, Movellan J (2005) Recognizing facial expression: machine learning and application to spontaneous behavior. IEEE Comput Soc Conf Comput Vis Pattern Recognit 2:568–573

Berretti S, Del Bimbo A, Pala P, Amor BB, Daoudi M (2010) A set of selected sift features for 3d facial expression recognition. In: International conference on pattern recognition, IEEE, pp 4125–4128

Cai J, Meng Z, Khan AS, Li Z, O'Reilly J, Tong Y (2018) Probabilistic attribute tree in convolutional neural networks for facial expression recognition. arXiv preprint arXiv:181207067

Cao Q, Shen L, Xie W, Parkhi OM, Zisserman A (2018) Vggface2: A dataset for recognising faces across pose and age. In: IEEE international conference on automatic face and gesture recognition, IEEE, pp 67–74

Chang FJ, Tran AT, Hassner T, Masi I, Nevatia R, Medioni G (2018) Expnet: Landmark-free, deep, 3d facial expressions. In: IEEE International conference on automatic face and gesture recognition, IEEE, pp 122–129

Chen T, Kornblith S, Norouzi M, Hinton G (2020) A simple framework for contrastive learning of visual representations. In: International conference on machine learning, PMLR, pp 1597–1607

Chen X, Pan Z, Wang P, Zhang L, Yuan J (2015) Eeg oscillations reflect task effects for the change detection in vocal emotion. Cogn Neurodyn 9(3):351–358

Deng Z, Choi KS, Jiang Y, Wang S (2014) Generalized hidden-mapping ridge regression, knowledge-leveraged inductive transfer learning for neural networks, fuzzy systems and kernel methods. IEEE Trans Cybern 44(12):2585–2599

Devlin J, Chang MW, Lee K, Toutanova K (2019) BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American chapter of the association for computational linguistics: human language technologies. 1:4171–4186

Fan X, Qureshi R, Shahid AR, Cao J, Yang L, Yan H (2020) Hybrid separable convolutional inception residual network for human facial expression recognition. In: International conference on machine learning and cybernetics, IEEE, pp 21–26

Feutry C, Piantanida P, Bengio Y, Duhamel P (2018) Learning anonymized representations with adversarial neural networks. arXiv preprint arXiv:180209386

He K, Zhang X, Ren S, Sun J (2015) Delving deep into rectifiers: Surpassing human-level performance on imagenet classification. In: IEEE International conference on computer vision, pp 1026–1034

He K, Zhang X, Ren S, Sun J (2016) Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition, pp 770–778

Hu P, Cai D, Wang S, Yao A, Chen Y (2017) Learning supervised scoring ensemble for emotion recognition in the wild. In: Proceedings of the 19th ACM international conference on multimodal interaction, pp 553–560

Kasiran Z, Yahya S (2007) Facial expression as an implicit customers' feedback and the challenges. IEEE

Khan S, Chen L, Zhe X, Yan H (2016) Feature selection based on co-clustering for effective facial expression recognition. Int Conf Mach Learn Cyberne 1:48–53

Khan S, Chen L, Yan H (2017) Co-clustering to reveal salient facial features for expression recognition. IEEE Trans Affect Comput 11:314

Khorrami P, Paine T, Huang T (2015) Do deep neural networks learn facial action units when doing expression recognition? In: IEEE International conference on computer vision workshops, pp 19–27

Kingma DP, Ba J (2015) Adam: a method for stochastic optimization. In: International conference on learning representations

Koujan MR, Alharbawee L, Giannakakis G, Pugeault N, Roussos A (2020) Real-time facial expression recognition" in the wild"by disentangling 3d expression from identity. In: International conference on automatic face and gesture recognition, IEEE

Krizhevsky A, Sutskever I, Hinton GE (2012) Imagenet classification with deep convolutional neural networks. Adv Neural Inf Process Syst 25:1097–1105

Langner O, Dotsch R, Bijlstra G, Wigboldus DH, Hawk ST, Van Knippenberg A (2010) Presentation and validation of the radboud faces database. Cogn Emot 24(8):1377–1388

Li M, Xu H, Huang X, Song Z, Liu X, Li X (2018) Facial expression recognition with identity and emotion joint learning. In: IEEE Transactions on affective computing

Li S, Deng W (2018) Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition. IEEE Trans Image Process 28(1):356–370

Li S, Deng W (2020) Deep facial expression recognition: a survey. IEEE Trans Affect Comput

Li S, Deng W, Du J (2017) Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild. In: IEEE Conference on computer vision and pattern recognition, pp 2852–2861

Lian Z, Li Y, Tao JH, Huang J, Niu MY (2020) Expression analysis based on face regions in read-world conditions. Int J Autom Comput 17(1):96–107

Lin TY, Dollár P, Girshick R, He K, Hariharan B, Belongie S (2017) Feature pyramid networks for object detection. In: IEEE conference on computer vision and pattern recognition, pp 2117–2125

Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I (2010) The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE conference on computer vision and pattern recognition workshops, IEEE, pp 94–101

Lundqvist D, Flykt A, Öhman A (1998) The karolinska directed emotional faces (kdef). Department of Clinical Neuroscience, Psychology section, Karolinska Institutet 91(630):2–2

Luong MT, Pham H, Manning CD (2015) Effective approaches to attention-based neural machine translation. In: Proceedings of the 2015 conference on empirical methods in natural language processing, pp 1412–1421

Mao Q, Rao Q, Yu Y, Dong M (2016) Hierarchical bayesian theme models for multipose facial expression recognition. IEEE Trans Multimedia 19(4):861–873

Mavani V, Raman S, Miyapuram KP (2017) Facial expression recognition using visual saliency and deep learning. In: IEEE international conference on computer vision, pp 2783–2788

Minaee S, Abdolrashidi A (2019) Deep-emotion: Facial expression recognition using attentional convolutional network. arXiv preprint arXiv:190201019

Mollahosseini A, Chan D, Mahoor MH (2016) Going deeper in facial expression recognition using deep neural networks. In: IEEE

Winter conference on applications of computer vision, IEEE, pp 1–10
Mollahosseini A, Hasani B, Mahoor MH (2017) Affectnet: a database for facial expression, valence, and arousal computing in the wild. IEEE Trans Affect Comput 10(1):18–31
Ocegueda O, Shah SK, Kakadiaris IA (2011) Which parts of the face give out your identity? In: IEEE conference on computer vision and pattern recognition, IEEE, pp 641–648
Pan SJ, Yang Q (2009) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359
Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L et al (2019) Pytorch: an imperative style, high-performance deep learning library. Adv Neural Inf Process Syst 32:8024–8035
Prieto LAB, Oplatkova ZK (2018) Emotion recognition using autoencoders and convolutional neural networks. Mendel 24(1):113–120
Ruiz-Garcia A, Elshaw M, Altahhan A, Palade V (2017) Stacked deep convolutional auto-encoders for emotion recognition from facial expressions. In: International joint conference on neural networks, IEEE, pp 1586–1593
Shahid AR, Khan S, Yan H (2020) Contour and region harmonic features for sub-local facial expression recognition. J Vis Commun Image Represent 73:102949
Shan C, Gong S, McOwan PW (2005) Robust facial expression recognition using local binary patterns. In: IEEE international conference on image processing, IEEE, 2:II–370
Shen F, Dai G, Lin G, Zhang J, Kong W, Zeng H (2020) Eeg-based emotion recognition using 4d convolutional recurrent neural network. Cogn Neurodyn 14(6):815–828
Shih FY, Chuang CF, Wang PS (2008) Performance comparisons of facial expression recognition in jaffe database. Int J Pattern Recognit Artif Intell 22(03):445–459
Sun W, Zhao H, Jin Z (2017) An efficient unconstrained facial expression recognition algorithm based on stack binarized auto-encoders and binarized neural networks. Neurocomputing 267:385–395
Szegedy C, Liu W, Jia Y, Sermanet P, Reed S, Anguelov D, Erhan D, Vanhoucke V, Rabinovich A (2015) Going deeper with convolutions. In: IEEE conference on computer vision and pattern recognition, pp 1–9
Szegedy C, Vanhoucke V, Ioffe S, Shlens J, Wojna Z (2016) Rethinking the inception architecture for computer vision. In: IEEE conference on computer vision and pattern recognition, pp 2818–2826
Szegedy C, Ioffe S, Vanhoucke V, Alemi AA (2017) Inception-v4, inception-resnet and the impact of residual connections on learning. In: Thirty-First AAAI conference on artificial intelligence
Tan C, Sun F, Kong T, Zhang W, Yang C, Liu C (2018) A survey on deep transfer learning. In: International conference on artificial neural networks, Springer, pp 270–279
Trepagnier CY, Sebrechts MM, Finkelmeyer A, Stewart W, Woodford J, Coleman M (2006) Simulating social interaction to address deficits of autistic spectrum disorder in children. Cyberpsychol Behav 9(2):213–217
Wang S, Liu Z, Lv S, Lv Y, Wu G, Peng P, Chen F, Wang X (2010) A natural visible and infrared facial expression database for expression recognition and emotion inference. IEEE Trans Multimedia 12(7):682–691
Wen G, Chang T, Li H, Jiang L (2020) Dynamic objectives learning for facial expression recognition. IEEE Trans Multimed 22:2914
Wen Y, Zhang K, Li Z, Qiao Y (2016) A discriminative feature learning approach for deep face recognition. In: European conference on computer vision, Springer, pp 499–515
Yaddaden Y, Adda M, Bouzouane A, Gaboury S, Bouchard B (2018) User action and facial expression recognition for error detection system in an ambient assisted environment. Expert Syst Appl 112:173–189
Yosinski J, Clune J, Bengio Y, Lipson H (2014) How transferable are features in deep neural networks? Adv Neural Inf Process Syst 27:3320–3328
Yu F, Koltun V (2016) Multi-scale context aggregation by dilated convolutions. In: International conference on learning representations
Zamir AR, Sax A, Shen W, Guibas LJ, Malik J, Savarese S (2018) Taskonomy: Disentangling task transfer learning. In: IEEE Conference on computer vision and pattern recognition, pp 3712–3722
Zavarez MV, Berriel RF, Oliveira-Santos T (2017) Cross-database facial expression recognition based on fine-tuned deep convolutional network. SIBGRAPI conference on graphics, Patterns and Images, IEEE, pp 405–412
Zeng H, Shu X, Wang Y, Wang Y, Zhang L, Pong TC, Qu H (2020) Emotioncues: emotion-oriented visual summarization of classroom videos. Trans Vis Comput Graph 27:3168
Zhang H, Su W, Yu J, Wang Z (2020) Identity-expression dual branch network for facial expression recognition. In: IEEE transactions on cognitive and developmental systems
Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503
Zhao H, Liu Q, Yang Y (2018) Transfer learning with ensemble of multiple feature representations. In: International conference on software engineering research management and applications, IEEE, pp 54–61