© 2020 Springer. Personal use of this material is permitted. Permission from Springer must be obtained for all other uses, in any current or future media, including reprinting/republishing this material for advertising or promotional purposes, creating new collective works, for resale or redistribution to servers or lists, or reuse of any copyrighted component of this work in other works.

Weakly Supervised Facial Expression Recognition Via Transferred DAL-CNN and Active Incremental Learning

Ying Xu[1], Jian Liu[1], Yikui Zhai*[1], Junying Gan[1], Junying Zeng[1], He Cao[1], Fabio Scotti[2], Vincenzo Piuri[2], and Ruggero Donida Labati[2]

[1]Department of Intelligent Manufacturing, Wuyi University, Jiangmen, China, 529020

[2] Dipartimento di Informatica, Universita' Degli Studi di Milano, via Celoria 18, 20133 Milano (MI), Italy

Corresponding author: yikuizhai@163.com

Abstract: In recent years, facial expression recognition (FER) has becoming a growing topic in computer vision with promising applications on virtual reality and human-robot interaction. Due to the influence of illumination, individual differences, attitude variation and etc., facial expression recognition with robust accuracy in complex environment is still an unsolved problem. Meanwhile, with the widely use of social communication, massive data is uploaded to the Internet, the effective utilization of those data is still a challenge due to noisy label phenomenon in the study of FER. To resolve the above-mentioned problems, Firstly, a double active layer (DAL) based CNN is established to recognize the facial expression with high accuracy by learning robust and discriminative features from the data, which could enhance the robustness of network. Secondly, an active incremental learning method was utilized to tackle the problem of using Internet data. During the training phase, a two-stage transfer learning method is explored to transfer the relative information from face recognition to FER task to alleviate the inadequate training data in deep convolution network. Besides, in order to make better use of facial expression data from website and further improve the FER accuracy, UFEDW (Unconstrained Facial Expression Database from Website) database is built in this paper. Extensive experiments performed on two public facial expression recognition database FER 2013 and SFEW 2.0 have demonstrated that the proposed scheme outperforms the state-of-the-art methods, which could achieve 67.08% and 51.90% respectively.

Keywords: Deep Convolutional Neural Network, Facial Expression Recognition, Two-Stage Transfer Learning, Weakly Supervised Active Incremental Learning

1 Introduction

Facial expression is a nonverbal manner, which could convey the emotion and message. In recent years, automatic facial expression recognition has becoming a hot topic in computer vision, which could be incorporated into a wide variety of applications, such as lie detection, surveillance, human-robot interaction, et al. Facial expression caused by facial muscle movements are subtle and abundant, the differences of the same category of expression is unbalanced, which makes facial expression recognition more difficult. Hence, how to capture and represent the robust features are the point of this paper.

Nowadays, Deep Convolutional Neural Networks (DCNN) [1] are playing more and more important role in artificial intelligence applications, such as fingerprint image recognition [2], speech recognition [3], medical image processing [4], action recognition[5], et, al. Contrary to traditional machine learning method which detect the facial key point and extract features manually, DCNN could discover the discriminative deep feature of raw images directly. In this paper, based on the characteristic of facial expression recognition, we construct a DAL-Net. DAL-Net contains a double active layer (DAL) and Softmax-MSE loss function. A double active layer includes Maxout and ReLU activation units, which could reduce the redundancy of deep features and improve network's expressiveness effectively. A new type loss function, Softmax-MSE, is adopted in DAL-Net, which fused the Softmax loss and MSE loss. These two losses complement each other and could optimize network training best.

As we all know, the adequate training of deep convolutional neural network requires abundant data, and the development of computing boosted the speed [6]. However, in real world, facial expression samples with label are scarce, and it takes high cost to obtain labeled samples by annotators. To solve this problem, many researchers have turned to utilizing transfer learning for further study [7-9]. Due to the exploration of the similarity of target domain and source domain, the transfer information of source domain could alleviate the data dependence of target domain. In this paper, we utilized a two-stage transfer learning method to take better usage of face recognition data information. Through dividing transfer learning into two stages, which freeze some layers in first stage and unfreeze these layers in second stage. Experimental results have demonstrated the effectiveness of two-stage transfer learning.

With the rapid development of technology, social software and smart devices, images in website with facial expression is huge，effective utilizing the facial expression images of website

could further improve the network's performance. In this paper, we construct W-FED database. Specifically, we utilized the four most used Internet search engines, which are Google, Baidu, Bing and Yahoo, to find facial expression images using expression related phrases as searching tags, to construct the W-FED database. However, due to the noisy label problem [10] with website images, W-FED database was restricted and can not be used directly for the training of network. In this paper, instead of abandoning the mislabeled samples, we adopted active incremental learning method to solve the noise label problem. After each round active selection, the most valuable samples in W-FED database were singled out. Then, those samples are viewed unlabeled and manually annotated. After the annotation, all samples were added into the network's training for incremental learning. The active incremental learning will stop until the DAL-Net convergence. As far as we know, this paper is the first to introduce active incremental learning into facial expression recognition.

The main contributions of this paper are as follows.

(1) A DAL-Net, which could reduce the redundancy of feature extraction and obtained the robust and discriminative feature extraction, is proposed to the facial expression recognition.

(2) A two-stage transfer learning method was utilized to transfer the relative information of face recognition to FER task, to alleviate the dependence of deep convolutional network on data volume.

(3) Construct a website facial expression database, W-FED database, to make full use of the facial expression images of website and further improve the recognition accuracy of FER.

(4) Combined the active selection and incremental learning, a noise label problem of W-FED database was solved.

The rest of this paper is organized as follows. Section 2 presents the related works. Section 3 describes the overall scheme. Section 4 analyzes experiments and compares the performance of the proposed method with other existing methods. Finally, Section 5 concludes this paper.

2 Related work

Facial expression recognition has always been an important research field in pattern recognition. Like many traditional pattern recognition tasks, FER research mainly focuses on feature extraction and classification. In feature extraction, there exists four methods. The first method is to extract geometric features from facial image, like the Active Shape Model (ASM)[11].

This method is conducted by locating facial parts like eyes, mouth, eyebrows and nose in facial images, getting their contour and facial structural map to obtain the geometric features. The second one is texture feature extraction, which focus on the color and pixel intensity distribution of facial expression images or the local area of images, such as the Local Binary Pattern (LBP)[12], the Local Phase Quantization (LPQ)[13] and the Histogram of Oriented Gridients (HOG)[14]. The third one is statistical feature extraction, which utilizes statistical method to calculate the information entropy and gray histogram of facial expression images, such as the Principal Component Analysis (PCA)[15], the Independent Component Analysis (ICA)[16], etc.. The last method is the transform domain feature-based extraction method, which the images were transformed from the geometric space into frequency space to describe the images, such as the Gabor Filters method [17].

For classification, there are four methods. The first one is the K-Nearest Neighbor (KNN)[18-19]. By calculating the similarity between the test image and the sample set images, the K most similar samples could be selected. Then, the largest category of K samples is considered to be the label of the test image. The second method is the Artificial Neural Networks (ANN)[20], which achieve image classification through a neural network. The third method is the Hidden Markov Model (HMM)[21], by constructing an category distribution to calculate the probability of which category the test image belongs to. The last method is the Support Vector Machine (SVM) [22-23], mapping facial expression features into high dimensional space and training a hyper-plane. The specific hyper-plane was used to classify the new image.

Compared with traditional feature expression recognition methods, deep learning method constructs a multi-layer nonlinear neural network that learns deep expression description from facial expression database. Liu et al. [24] proposed an LBP/VAR feature for facial expression recognition. LBP/VAR is robust for rotation and illumination, which was sent to a deep belief network (DBN) to extract deep facial expression features, could achieve satisfied recognition result. Liu et al. [25] constructed a deep convolutional neural network for facial expression recognition, which had receptive field construction and group-wise block, and combined the local appearance feature to optimize the recognition result.

Recently, the rapid development of deep learning has brought a new sight to the FER task. Lecun et al. [26] proposed DCNN, which is the first successful learning algorithm for multi-layer

networks. Though local spatial mapping and sharing weights, DCNN could reduce the parameter calculation of back propagation and speed up the network's performance. Then, many researchers adapted the DCNN framework in their FER research and achieved satisfied performances. Burkert et al. [27] instead extracted the traditional manual feature, they adopted a novel DCNN framework to automatically achieve the facial expression recognition. Yu et al. [28] proposed a DCNN FER framework to fused three state-of-the-art face detectors, which achieved satisfactory performance on the SFEW 2.0 dataset.

The adequate training of DCNN needs large amount of training samples, but the sample amount is inadequate in real FER task. For facial expression recognition task, the sample amount of public database is deficient, it tends to lead overfitting phenomenon. To tackle the data deficient problem for such a small database task, the utilization of transfer learning strategy before the training of the target domain maybe a good solution. Zhang et al. [29] used two transfer learning method, multi-kernel learning and multi-task learning, to further learn the facial expression feature characteristic. Chen et al. [30] transferred the information of face recognition to FER task, which achieved some extend improvement of facial expression recognition. Besides, there are also some researchers utilized transfer learning method in FER task [31-32].

With the development of social media, massive facial expression images are available from Website. Due to the noise label problem of Website images, however, the facial expression images could not be directly used for the training of deep convolutional neural network. For this thorny issue, Yu et al. [33] proposed SVM based active learning method to select the fine quality facial expression images of Website, which without a suitable evaluation criterion to measure the worthiness of Website images. Sukhbaatar et al. [34] introduced a noisy label layer into the deep convolutional neural network, which could help the network to learn the distribution of noisy samples.

3 Proposed Approach

In this paper, the various parts of the work will be introduced. Firstly, an effective convolution neural network using a double activation function and Softmax-MSE loss function is used for facial expression recognition task. Moreover, a two-stage transfer learning method was utilized to alleviate the overfitting phenomenon, which transfer the pre-trained information of large-scale face recognition database CASIA-WebFace to FER task. Finally, a weakly supervised

active incremental learning method was adopted for select the most valuable samples of UFEDW database, then added these valuable samples iteratively to active incremental the training of the proposed net for FER task. The whole proposed approach of this paper is shown in Fig.1.

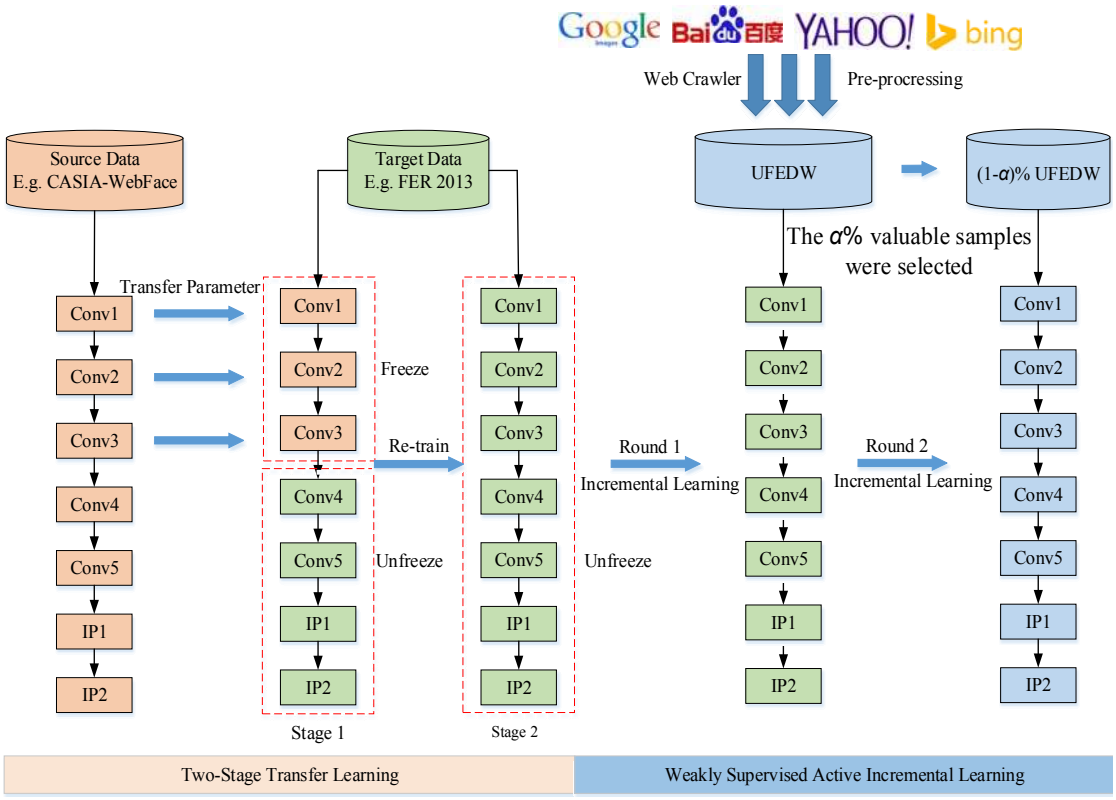


Fig.1 The whole scheme of this paper

3.1 Proposed Network Architecture

The proposed convolution neural network architecture, as shown in Fig.2, including five convolution layers, two fully connection layers, a double activation function and a Softmax-MSE loss layer. The specific parameter setting is shown in Table 1. In Fig.2, “Conv” presents convolution layer, “MFM” presents Max Feature Map (MFM) layer, “Pool” presents Pooling layer, “IP1” and “IP2” are fully connection layers and “Last Conv” means the last convolution layer in the proposed net. The first convolution layer’s kernel is 5, and output 96 feature maps for the next layer. Then, a MFM layer was used to processed these feature maps, which introduce the non-linear factors to the network. In addition, a double activation layer is added before the fully connection layer, and the complementary advantages of Maxout and ReLU active function were utilized to speed up the proposed net converge to the global optimal solution. Finally, the fully

connection layer is connected to the Softmax classifier and Softmax-MSE loss layer, which was used to avoid the over-fitting in the model training process.

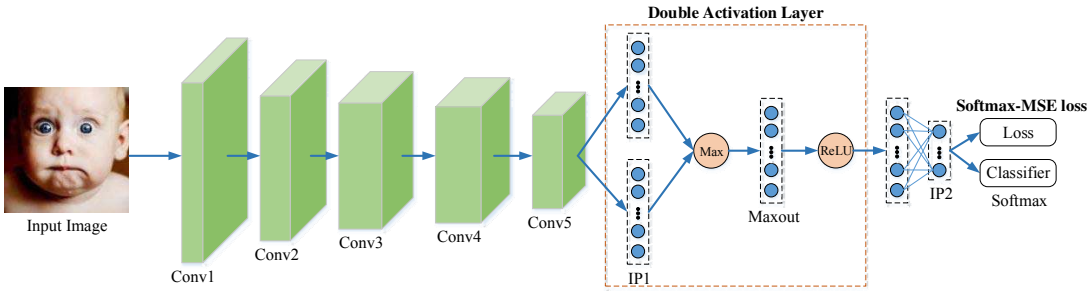


Fig.2 The architecture of the proposed net

Table 1 The parameter setting of the proposed net

Layer	Parameters
Input	144×144
Conv1	96, 5×5
Conv2	192, 3×3
Conv3	384, 3×3
Conv4	384, 3×3
Conv5	256, 3×3
IP1	512
IP2	7
Loss	Softmax-MSE

3.1.1 Double Activation Layer

The activation layer is an indispensable part of CNN, which could introduce the nonlinear mapping and improve the nonlinear fitting ability of the model. Traditional activation functions, such as Sigmoid and Tanh, are widely used cause of their strong nonlinear transformation ability. However, this type activation functions are also resented by researches for their vulnerability to vanishing gradient, which lead the training down.

ReLU is a piecewise rectified linear activation function, which output the input directly if is positive, otherwise, it will output zero. This simple and crude manner of forcing some data to be zero has been proved by practice that the network after training is moderately sparse. Due to the sparse characteristic, the ReLU activation function could speed up the converge of SGD than Sigmoid and Tanh, and then it replaces the traditional activation function.

Maxout activation function [35] divided the input maps into two parts, the counterpart's neurons of these two parts will be compared, then the max part will be preserved. Maxout

activation function could fitting into any convex function, and reduce the gradient decline in the process of network's training. Hence, the Maxout activation function could speed up the network convergence to a global solution.

From the above analysis, to make full use of their advantages, this paper presents a novel double activation layer (DAL), which could speed up the proposed net to converge to its global optimal solution. The architecture of DAL is shown in Fig.3.

Specifically, the output feature maps of the "Last Conv" layer was divided into two groups, through the Maxout architecture, the max counterparts were remained. Then, the max counterparts were pumped into the ReLU architecture.

Assuming that input of Maxout layer is X, and X is divided into parts X_1 and X_2, then the output of Maxout is

$$Y = \max(X_1, X_2) \tag{1}$$

Supposing that the input of ReLU layer is X, then the output of ReLU layer is

$$Y = \max(0, X) \tag{2}$$

After the Maxout function, the output of the ReLU layer is connected to IP2 layer. The IP2 layer, which has denoted by "7", is used as the input of the Softmax-MSE loss function.

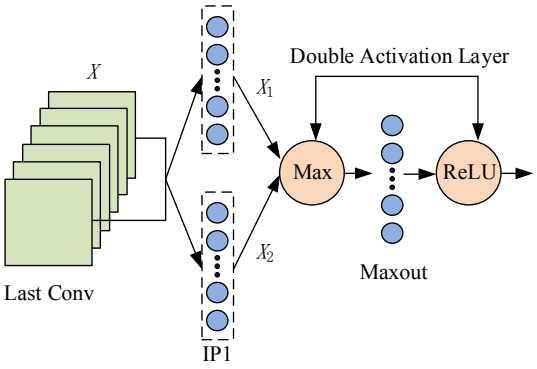


Fig.3 Architecture of double activation layer

3.1.2 Softmax-MSE Loss Function

There are mainly two issues transferring the face recognition net into the facial expression net. The first one is that the facial expression net may contain redundant information which is helpful for face recognition because of the big quantitative difference between face and facial expression database. The second problem is that the fine-tuned face recognition network is so huge for FER task that the over-fitting problem can not be solved appropriately.

To overcome the above problems, Softmax-MSE loss function is proposed. Different from the non-cost-sensitive loss function Softmax, Softmax-MSE is cost-sensitive which takes classification error into consideration. Experimental results indicate that Softmax loss function would achieve better performance in some tasks such as face recognition which do not care the correlations between the prediction. For cost-sensitive function like Euclidean loss function would achieve better regressive performance on tasks that require high correlation of prediction. The Softmax-MSE loss function adopted in this paper, combines the advantages of this two-loss function and complements each other. The detail of Softmax-MSE loss function as follows.

For m category recognition task, the input of Softmax layer is $X = \{x_0, x_1, \cdots, x_{m-1}\}$. The calculation of Softmax loss function is

$$p_k = \frac{e^{x_k - \max(X)}}{\sum_{i=0}^{m-1} e^{x_i - \max(X)}} \tag{3}$$

where $k \in [0, m-1]$, p_k presents the probability of the sample belongs to the k-th category. If the input batch size of the network is n, $p_k = \max([p_0, p_1, \cdots, p_{m-1}])$, then the regressive prediction value is

$$\hat{y}_j = \sum_{k=0}^{m-1} k p_k \tag{4}$$

The Softmax-MSE loss value of the nth image is

$$L = \frac{1}{n} \sum_{j=0}^{n-1} (\hat{y}_j - y_j)^2 \tag{5}$$

where y_j is the ground truth label of jth image, $\hat{y}_j$ is the prediction label of jth image. In order to maintain the advantage of Softmax loss function, the gradient of Softmax loss function was used in Softmax-MSE layer is

$$\frac{\partial L_j}{\partial x_i} = \begin{cases} p_i - 1, & i = y_j \\ p_i, & i \neq y_j \end{cases} \tag{6}$$

3.2 Two-Stage Transfer Learning

Due to the lack of labeled facial expression samples, the performance of proposed net was restricted. Transfer learning is a method, utilizing the information of other cross-filed task to enhance the training of the target task. Due to transfer more relative information for the network's training, the recognition performance of target task achieves some extent improvement. For small

sample task, transfer learning could alleviate the dependency of deep convolution network on labeled data amount. Different from the traditional machine learning method, which focuses on training different model for different tasks, transfer learning method focuses on transfer the

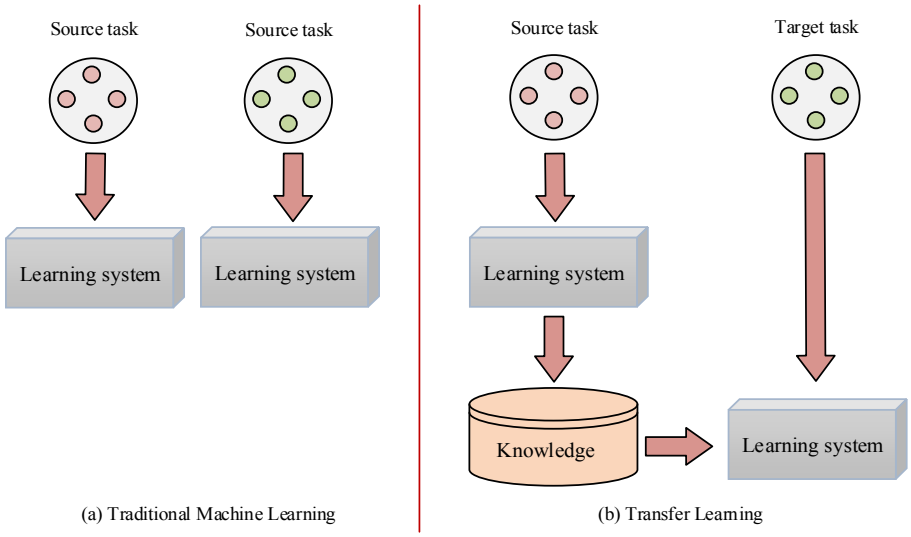


Fig.4 Traditional machine learning process and transfer learning process

knowledge of source domain to target domain in a common model. The difference between traditional machine learning method and transfer learning method is illustrated in Fig.4.

Transfer learning [36] could be divided into three types, inductive transfer learning, transductive transfer learning and unsupervised transfer learning. The transfer learning method that has the same source and target domain and irrelevant source and target task is classified as an inductive transfer learning method. The transfer learning method that shares the same domain but differs in source and target task is defined as transductive transfer learning. The transfer learning method that differs in source and target domain as well as in source task and target task is defined as unsupervised transfer learning. In this paper, an inductive transfer method, two-stage transfer learning, was utilized to transfer the relative information of face recognition to FER task, to alleviate the dependence of deep convolution network on data volume.

A. Definition of Transfer Learning

Given a source domain D_S and a target domain D_T, which corresponding to learning task T_S and T_r. The purpose of transfer learning is to improve the learning ability of target prediction function $f(T(\cdot))$ in D_T using the knowledge of D_S and T_S, where $D_S \neq D_T$, or $T_s \neq T_T$.

More specific, for facial expression recognition task, the source domain is defined as $D=\{F, P(X)\}$, where $F=\{f_1, f_2, ..., f_n\}$ is a feature space with n dimensions, f_i is a feature, $X=\{x_1, x_2, ..., x_n\}$

is facial expression database, and $P(X)$ is marginal probability distribution of X. For a domain that is thought to be different, the feature spaces or marginal probability distribution is different. The task domain is defined as $T=\{y, P(y|X)\}$, where y is the label space and $P(y|X)$ is the classification model.

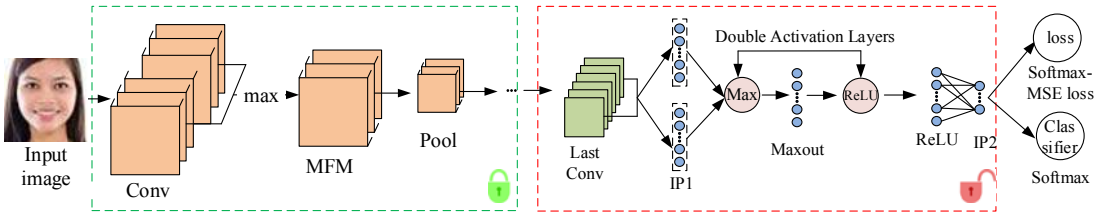
B. Two-stage Transfer Learning

In this paper, the source task is face recognition, the target task is facial expression recognition. Since face recognition and facial expression recognition tasks have different marginal probability distribution and the same feature space, the proposed net's performance could be improved by transfer learning. Specifically, in this paper, we pre-trained the proposed net for face recognition task by CASIA-WebFace database at first. Then the concrete implementation process of two-stage transfer learning is as follows.

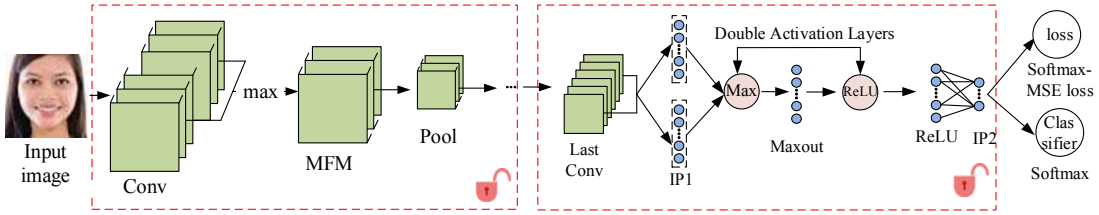
Stage 1: Firstly, after transfer the parameter of proposed net, few bottom convolution layers were frozen and the parameter update was closed, only the layers below it update the parameter. Then, the proposed net starts training on the facial expression database until the network converges to the optimum. In the end, the optimal model will used for Stage 2. In Stage 1, the related information of face recognition is transferred to the facial expression recognition task, which improve the target task performance.

Stage 2: In this Stage, the convolution layers frozen in the Stage 1 were unfrozen, and the whole network will be re-trained based on the current optimal model. A smaller learning rate than Stage 1 was adopted for the training of Stage 2. This Stage could fine-tune the whole network parameters slightly, to make the model more suitable for the facial expression recognition task. Experimental results show that the two-stage transfer learning method alleviate the overfitting phenomenon and improve the performance of facial expression recognition in some extent.

The diagram of two-stage transfer learning is shown in Fig.5. As shown in Fig.5(a), we freeze the parameter update of the few bottom convolution layers. Specifically, the parameter learning rate "lr_mult" of these frozen layers is set to 0. The other layers set parameter learning rate as 1, which will update the information in the Stage 1. In Stage 2, all the parameters update was unfrozen, and the overall re-training was carried out in proposed net. Two-stage transfer learning could make better use of the information which transferred from the face recognition task, and the re-trained proposed net could make the model more suitable for facial expression recognition task.



(a) Training Stage 1



(b) Training Stage 2

Fig.5 Two-stage transfer learning

3.3 Weakly Supervised Active Incremental Learning

Supervised learning is an effective and intuitive way of training the CNN, but requires the annotation of every sample in the training database. To reduce the dependency of the labeled data, deep learning models can work under weakly supervised schemes. Active incremental learning method, which is a subclass of weakly supervised learning method, uses data samples that may contain annotation noise. Through an expert iteratively annotates data, active incremental learning method could maximize network performance with minimum the cost of labeling data.

To make full use of website images, in this paper, a UFEDW database was built. However, the images of UFEDW database are very diverse and different, some images either contain mismatched labels or even contains no realistic human face in the images. It is not realistic to carry out manual label correction and filtering the whole images of UFEDW database directly, cause the time and labour cost increased greatly. To minimizing the cost of manual label correction and filtering, in this section, we adopted weakly supervised active incremental learning method to selected the most valuable samples from the UFEDW database iteratively and stepped improve the training of the proposed net. The diagram of weakly supervised active incremental learning algorithm is shown in Alg.1. Among them, Z presents the newly selected candidates, H presents the misclassified candidates. Assuming the prediction of patch x_i^j by the current CNN is p_i^j, the histogram of candidate C_i is $P_i = p_i^j, j \in [1, m]$.

In weakly supervised active incremental learning, the key is to develop a criterion for determining the "worthy" of a candidate for annotation. Due to fact that, all patches from the same candidate shares the same label, are naturally expected to have similar predictions by the current optimal model. To evaluate the consistency of multiple patches in one candidate, those patches entropy and diversity were calculated. If the ground truth label of these patches is same as the current optimal model's prediction label, annotated this candidate is worthless cause it will have no lead performance improvement.

In this paper, a criterion that combines entropy and diversity was utilized to calculate the "worthiness" of the UFEDW database image. Entropy measures the certainty of the facial expression recognition, images with higher entropy denotes a higher degree of information. Diversity measures the predictive consistency among patches within image. Mathematically, we present the calculation of entropy and diversity as follows.

Suppose that $U = \{C_1, C_2, \ldots, C_n\}$, where C_i is candidate image in UFEDW database. For each candidate, $C_i = \{x_i^1, x_i^2, \ldots, x_i^j\}$, where x_i^j is one patch of the image which is divided into j patches. The entropy of each candidate image C_i is defined as

$$e_i = -\frac{1}{m}\sum_{k=1}^{|\gamma|}\sum_{j=1}^{m} p_i^{j,k} \log p_i^{j,k} \tag{7}$$

where $|\gamma|$ presents the prediction label of this candidate image, and $p_i^{j,k}$ is the recognition probability of one image.

The diversity of each candidate image C_i is defined as

$$d_i = \sum_{k=1}^{|\gamma|}\sum_{j=1}^{m}\sum_{l=j}^{m}(p_i^{j,k} - p_i^{l,k})\log\frac{p_i^{j,k}}{p_i^{l,k}} \tag{8}$$

Then, the "worthiness" of each candidate C_i is defined as

$$W_i = \lambda_1 e_i + \lambda_2 d_i \tag{9}$$

where λ_1 and λ_2 are the trade-off between entropy and diversity. During experiments on the UFEDW database, λ_1 and λ_2 were set as 1 and 0.5, respectively.

For example, an image is shown as Fig.6(a), the original image was cropped into 9 patches which is shown as Fig.6(b). The predicted confidence of each patches is shown in Fig.5(c), it's obvious that not all the patches are informative. To enhance the robustness and reduce the computation of

our method, we descending the patches confidence, and the top α% patches were used to calculate the "worthiness".

Alg.1: Weakly supervised active incremental learning method

Input:

$U = \{C_i\}, i \in [1,n]$ {U presents UFEDW database, which contains n candidates}

$C_i = \{x_i^j\}, j \in [1,m]$ {m presents the patch number of C_i}

M_0 presents the optimal model after two-stage transfer learning; α presents the selection ratio of patches; b presents batch size; γ presents the category of facial expression recognition.

Output:

L presents labeled candidates; M_T presents the fine-tuned model at round T

$L \leftarrow \phi$; $T \leftarrow 1$

repeat

for each $C_i \in U$ **do**

$P_i \leftarrow M_{T-1}(C_i)$ {outputs of M_{T-1} given $\forall x \in C_i$}

$C_i' \leftarrow C_i$ descending order according to the predicted dominant class

$\hat{y} \leftarrow arg\,max_{y \in \gamma} \sum P_i^y$

$C_i^\alpha \leftarrow$ top $\alpha \times 100\%$ of the patches of the sorted list C_i'

Compute W_i for C_i^α (Eq. 9)

end

Sort U according to W in descending order

Compute sampling probability W^s using sorted list W' (Eq.11 and Eq.12)

Associate labels for b candidates with sampling probabilities: $Z \leftarrow Q(W^s, b)$

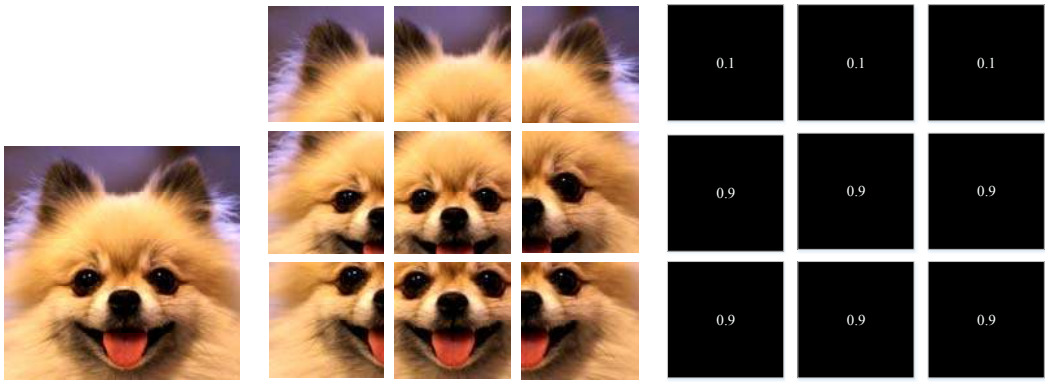
$P \leftarrow M_{T-1}(L)$ {outputs of M_{T-1} given $\forall x \in L$}

Select misclassified candidates from L based on their annotation : $H \leftarrow J(P, L)$

fine-tune M_{T-1} with $H \bigcup Z$: $M_T \leftarrow F(H, M_{T-1})$

$L \leftarrow L \bigcup Z ; U \leftarrow U \setminus Z ; T \leftarrow T+1$

until classification performance is satisfied.



(a) The original image (b) The cropped image patches of the original image (c) The predictive value of image patches generated by model

Fig.6 The example of image's patches

Then the dominant category $\hat{y}$ of this candidate C_i is calculated ,which is defined as the category with the highest confidence in the mean prediction on the current model, that is

$$\hat{y} = \underset{y \in \gamma}{arg\,max} \sum P_i^y \tag{10}$$

where P_i^y is the confidence output of label y. After sorting P_i according to $\hat{y}$, Eq.9 was used to calculate the "worthiness" of top α% patches.

Inspired by Zhou [37], the random selection may superior active incremental learning at the beginning, because the active incremental learning depends on the current optimal model to select samples for annotation. As a result, the poor selection degrades the performance of subsequent selections. Hence, in this paper, inject the random selection into active incremental learning. Specifically, we inject randomization in our method by selecting actively according to its sampling probability W_i^s . The calculation is shown as below.

$$W_i^{'} \leftarrow (W_i^{'} - W_{wb}^{'}) / (W_1^{'} - W_{wb}^{'}), \forall i \in [1, wb] \tag{11}$$

$$W_i^s \leftarrow W_i^{'} / \sum_i W_i^{'}, \forall i \in [1, wb] \tag{12}$$

where $W_i^{'}$ is descending sorted W_i, w presents random extension. Assuming the b number of candidates are required for annotation. Instead of selecting top b candidates, we extend candidates selection pool to $w \times b$, and sample the candidates from W_i^s .

4 Facial Expression Database

Currently, the open-source Facial Expression Database is small scale relatively, the DCNN can not training sufficiently. With the rapid development of smart devices, it's easier to share and spread their daily photos on the Internet. This huge photo pool could give the abundant facial expression images to further improve the training of the proposed net. To make better use of facial expression images in Internet, we adopted four search engines to downloaded the counterpart images though the facial expression tags. After pre-processing these images, the UFEDW database was constructed.

4.1 Database Construction

Specifically, we used Google, Baidu, Bing and Yahoo as the image search engine. Through the searching tags, "Angry", "Disgust", "Fear", "Happy", "Sad", "Surprise" and "Neutral", the web crawler downloaded the counterpart images. After the image crawling, 51387 images were downloaded from the website.

Then, the Viola Jones [38] face detector was applied here for image cropping and automatic database cleaning. After detecting the five facial landmarks by the face detector, only 32,713 images remained. These images ware cropped and aligned into a fixed size of 144×144 gray images, then the UFEDW database is built. The pre-processing diagram of the UFEDW database is shown in Fig.7.

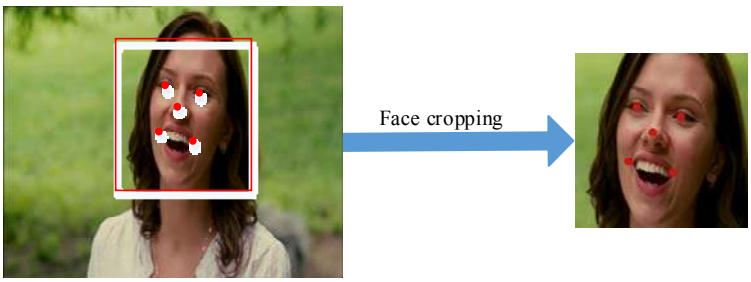


Fig. 7 The pre-processing diagram of UFEDW database

4.2 Sample Images and Data Distribution

The samples of UFEDW are from the Internet, the diversity of collection equipment leads to the varies of image quality, including a wide range of changes in the face image, such as age, expression, angle, illumination, and etc. For better presentation of the UFEDW database, some samples of database are shown in Fig.8. As can be seen obviously from the Fig.8, the data in the UFEDW database are very diverse and different. Besides this, the UFEDW faces the noise label

problem, and these noisy images either contain mismatched labels or even contains no realistic human face in the images. The noise label problem will restrict the use of the UFEDW database for the training of the proposed net. However, it is not realistic to carry out manual label correction and filtering the 32,713 images directly, cause the time and labour cost increased greatly. Besides that, it is not necessary to correction all the images of UFEDW, the most valuable images of it could assist the proposed net obtain the optimal resolution. In this paper, a weakly supervised active incremental learning was utilized to select the most valuable samples of UFEDW database iteratively, through the minimal samples achieve the maximize performance improvement. In addition, the distribution of the 7-class images in UFEDW, FER 2013 and SFEW 2.0 database are shown in Fig.9.

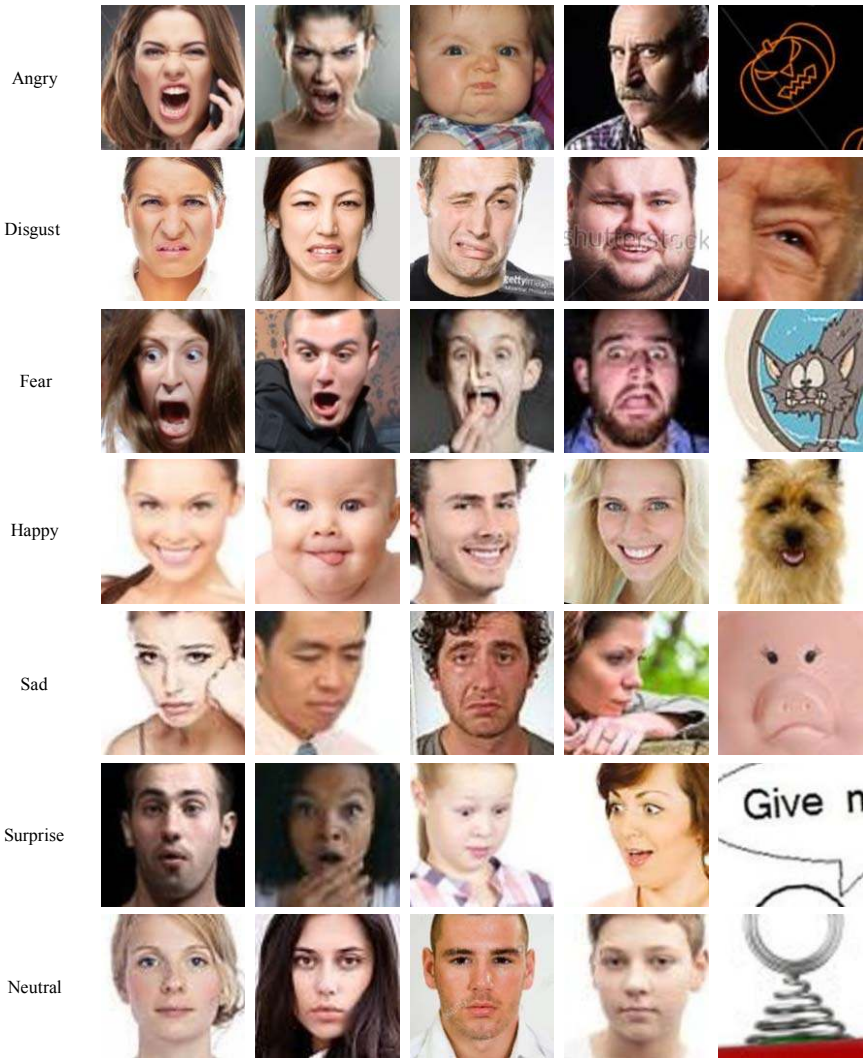


Fig. 8 Examples of W-FED database

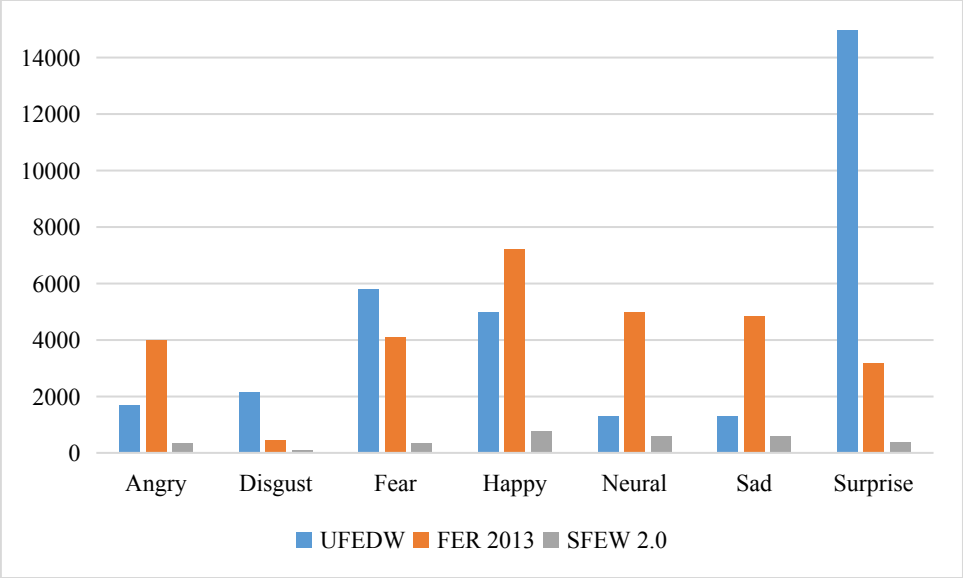


Fig. 9 Data distribution of UFEDW, FER 2013 and SFEW 2.0 database

4. Experiments and Analysis

In this paper, the experimental were conducted in three databases, SFEW2.0, FER2013 and W-FED, which specific information is shown in Table 2. Among them, SFEW2.0 is an enhanced version of SFEW database. The SFEW has only contains 958 training data, the data volume is insufficient for the training of deep convolutional network. Therefore, the data enhancement is realized by adding random noise to the three-channel gray image, respectively. At the same time, the sample of SFEW database which could not detect the face were removed automatically.

Table 2. Specific information of SFEW2.0, FER2013 and W-FED database

Database	Angry	Disgust	Fear	Happy	Neural	Sad	Surprise
SFEW2.0	356	80	327	773	580	593	385
FER2013	3995	436	4097	7215	4965	4830	3171
W-FED	1673	2146	5793	4973	1277	1277	15574

4.1. Experiments on two-stage transfer learning

To explore the effective of the two-stage transfer learning method, the experiments on the SFEW2.0 and FER2013 database were performed in this section. In the meantime, to specific observe the performance improvement of the facial expression recognition by the two-stage transfer learning method, we performed transferring parameters on all convolution layers of the DAL-Net separately. Table 3 and Table 4 shows the specific results of transfer the different convolution layers of DAL-Net in two stage on SFEW2.0 and FER2013, respectively. Among

them, the frozen layers set 3 means bottom layers to the third convolution layers were frozen in Stage 1, and all layers in DAL-Net were unfrozen in Stage 2. In particular, the frozen layers set 0 means all the parameter of face recognition task were transferred into DAL-Net for facial expression recognition.

Table 3. Recognition accuracy of two-stage transfer learning on FER2013 database

Experiment Number	Frozen layers	Stage 1 accuracy	Stage 2 accuracy
1	0	61.50%	64.70%
2	1	62.28%	64.45%
3	2	62.44%	64.76%
4	**3**	**62.30%**	**65.04%**
5	4	60.74%	63.76%
6	5	58.54%	63.87%

Table 4. Recognition accuracy of two-stage transfer learning on SFEW2.0 database

Experiment Number	Frozen layers	Stage 1 accuracy	Stage 2 accuracy
1	0	46.49%	46.67%
2	1	46.25%	46.01%
3	2	44.35%	44.17%
4	3	44.52%	44.94%
5	**4**	**47.68%**	**48.21%**
6	5	46.01%	47.26%

In Table 3, for FER2013 database, the overall accuracy in training Stage 2 are higher than Stage 1. We can see from experiment 1 to experiment 6 that overall accuracy increases and then decreases gradually, while the model in experiment 3 and the model in experiment 4 could achieve roughly the same performance in training Stage 1. While in the training Stage 2, the model in experiment 4 achieved the highest performance.

In Table 4, for SFEW2.0 database, the overall accuracy in the training Stage 2 are also higher than Stage 1. The model in experiment 5 achieved the highest performance during the Stage 2. With the further increase number of frozen layers, the performance of DAL-Net degrades. These two experiments show that, comparing conducted the transfer learning in a total when frozen layers set to 0, the transfer learning divided into two stage could achieve better performance improvement, validated the effectiveness of two-stages transfer learning. Furthermore, for two stage transfer learning, the confusion matrix of DAL-Net on FER2013 and SFEW2.0 were shown in Table 5-8, to show the facial expression recognition effect more clearly.

Table.5 Confusion matrix of DAL-Net on FER2013 database in training Stage 1

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.55	0.00	0.07	0.06	0.12	0.18	0.02
Disgust	0.52	0.09	0.13	0.04	0.05	0.18	0.00
Fear	0.13	0.00	0.29	0.06	0.12	0.28	0.12
Happy	0.02	0.00	0.01	0.86	0.06	0.03	0.03
Neutral	0.06	0.00	0.03	0.10	0.61	0.19	0.01
Sad	0.10	0.00	0.07	0.05	0.20	0.56	0.01
Surprise	0.02	0.00	0.08	0.06	0.03	0.02	0.78

Table.6 Confusion matrix of DAL-Net on FER2013 database in training Stage 2

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.58	0.00	0.07	0.06	0.10	0.17	0.02
Disgust	0.46	0.25	0.13	0.04	0.07	0.05	0.00
Fear	0.14	0.00	0.34	0.03	0.13	0.26	0.10
Happy	0.02	0.00	0.01	0.85	0.07	0.02	0.03
Neutral	0.05	0.00	0.03	0.08	0.67	0.16	0.01
Sad	0.11	0.00	0.08	0.04	0.19	0.57	0.01
Surprise	0.02	0.00	0.07	0.05	0.03	0.02	0.81

Table.7 Confusion matrix of DAL-Net on SFEW2.0 database in training Stage 1

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.36	0.04	0.00	0.09	0.23	0.06	0.21
Disgust	0.05	0.10	0.00	0.20	0.42	0.14	0.10
Fear	0.09	0.02	0.09	0.16	0.27	0.13	0.24
Happy	0.02	0.00	0.01	0.84	0.10	0.03	0.01
Neutral	0.00	0.00	0.07	0.09	0.63	0.10	0.12
Sad	0.04	0.02	0.04	0.13	0.13	0.56	0.08
Surprise	0.04	0.05	0.04	0.12	0.36	0.14	0.25

Table.8 Confusion matrix of DAL-Net on SFEW2.0 database in training Stage 2

	Angry	Disgust	Fear	Happy	Neutral	Sad	Surprise
Angry	0.39	0.08	0.00	0.12	0.17	0.04	0.20
Disgust	0.05	0.10	0.00	0.21	0.39	0.16	0.10
Fear	0.17	0.02	0.08	0.18	0.26	0.11	0.19
Happy	0.02	0.00	0.02	0.85	0.06	0.05	0.00
Neutral	0.00	0.00	0.06	0.11	0.64	0.09	0.11
Sad	0.02	0.02	0.04	0.13	0.16	0.56	0.08
Surprise	0.03	0.05	0.04	0.12	0.38	0.14	0.24

4.2. Experiments on DAL-Net architecture

To validate the effectiveness of DAL and Softmax-MSE structure, comparative experiments with or without these architectures were conducted. The recognition results are shown in Table 9. From Table 9, in FER2013 database, after utilizing the DAL and Softmax-MSE structure, the network's overall recognition accuracy improved 1.7% to 1.95%. For SFEW2.0 database, utilizing the DAL and Softmax-MSE structure could achieve 0.83% to 2.56%, which validate the effect of DAL-Net architecture.

Table.9 Recognition accuracy of the DAL-Net with or without DAL and Softmax-MSE

Database	Network	Stage1	Stage2
FER2013	DAL-Net (without DAL and Softmax-MSE)	60.35%	63.34%
	DAL-Net	62.30%	65.04%
SFEW2.0	DAL-Net (without DAL and Softmax-MSE)	45.12%	47.38%
	DAL-Net	47.68%	48.21%

4.3. Experiments on active incremental learning

In order to verify the validation of active incremental learning method, the specific experimental results on FER2013 and SFEW2.0 database were recorded in Table 10. Based on the current optimal pre-trained model of two-stage learning, α% images of W-FED database, which are most worthy for improve the network's performance, were selected. In experiment, we selected the top 2000 worthy images from W-FED database in each round, and correct their labels. Then, these samples were utilized for fine-tuning the current optimal DAL-Net model to further increase network's training. Only after two rounds, the performance of DAL-Net reached the optimal result. The active incremental learning could not only reduce the cost of annotated the noise sample, but also optimize the network's recognition performance progressively.

Table.10 Experiment results of DAL-Net using active incremental learning method

Database	Active incremental learning	Stage1	Stage2
FER2013	Original	65.01%	65.49%
	Round 1	66.13%	66.57%
	Round 2	67.02%	**67.08%**
SFEW2.0	Original	48.15%	48.15%
	Round 1	51.25%	52.02%
	Round 2	51.90%	**51.90%**

In Table 10, Original presents the situation, which the optimal DAL-Net model only utilized the two-stage learning method. It is obviously that, after the active incremental learning, the overall accuracy in both Stages have been boosted than original. This phenomenon shows that, the small number but the worthiest samples could further improve the recognition performance of DAL-Net. To solve the noise label problem of website images, instead of all the images on W-FED database were used for the training of DAL-Net, only the most valuable samples were selected. This strategy could not only save the computation time but also the labour to annotated the facial expression images.

4.4. Performance comparison with other methods

In order to verify the effectiveness of the whole scheme proposed in this paper, performance comparison experiment with other methods was conducted on the FER2013 and SFEW2.0 database. The recognition accuracy is shown in the Table 11.

Table 11. Performance comparison with other methods

Database	Methods	Recognition Accuracy
SFEW2.0	Inception [39]	47.70%
	Mapped LBP [40]	41.92%
	VGG	39.55%
	VGG with transfer learning	41.23%
	DAL-Net (ours)	**48.21%**
	DAL-Net+active incremental learning(ours)	**51.90%**
FER2013	Transfer Learning [41]	48.50%
	CNN [28]	57.10%
	DAL-Net (ours)	**65.04%**
	DAL-Net+active incremental learning(ours)	**67.08%**

As can be seen in Table 11, the whole scheme proposed in this paper could achieve better performance than other traditional methods. Besides that, comparing with other deep learning methods, the proposed scheme could achieve the state-of-the-art performance.

5. Conclusion

Facial expression recognition in the wild is a challenge task, due to the various factors degrades it's performance. In this paper, we present a novel scheme for facial expression recognition task to improve its real time performance. Firstly, a DAL-Net, which contains the double activation layer and Softmax-MSE loss function, is proposed to learn the robust and

discriminative deep features from the data. Then, a two-stage transfer learning method was utilized for alleviate the overfitting phenomenon, which caused by the insufficient amount of training data. To make full use of the website images, a W-FED database was established. Finally, to solve the noise label problem of W-FED database, an active incremental learning method was used to select the worthiest mislabeled samples to further improve the training of DAL-Net. Experimental results on two public facial expression databases, FER2013 and SFEW2.0, shows the advantages of the proposed scheme.

Compliance with Ethical Standards:

This work is supported by National Natural Science Foundation of China (No.61771347), Characteristic Innovation Project of Guangdong Province (No.2017KTSCX181); Young Innovative Talents Project of Guangdong Province(2017KQNCX206); Jiangmen Science and Technology Project ([2017] No.268); 2017 Guangdong Science and Technology Plan Project (No. 2017A010101019); Youth Foundation of Wuyi University (No.2015zk11); the Opening Project of GuangDong Province Key Laboratory of Information Security Technology(Grant No. 2017B030314131); the 2018 Opening Project of GuangDong Province Key Laboratory of Digital Signal and Image Processing.

Conflict of Interest: All Authors declares that he/she has no conflict of interest.

Ethical approval: This article does not contain any studies with human participants or animals performed by any of the authors.

References

[1] Zhou Y, Xu T, Zheng W, et al. Classification and recognition approaches of tomato main organs based on DCNN [J]. Transactions of the Chinese Society of Agricultural Engineering, 2017,33(15):219-226.

[2] Impact of digital fingerprint image quality on the fingerprint recognition accuracy [J]. Multimedia Tools and Applications, 78(3), 3649-3688.

[3] The optimization of speech recognition based on convolutional neural network [C]. IJHPCN 13.2 (2019): 222-231.

[4] Efficient quantum information hiding for remote medical image sharing [J]. IEEE Access, 6,

21075-21083.

[5] Real-time human action recognition using depth motion maps and convolutional neural networks [J]. International Journal of High Performance Computing and Networking 13.3 (2019): 312-320.

[6] Performance identification in large-scale class data from advanced facets of computational intelligence and soft computing techniques [J]. International Journal of High Performance Computing and Networking 13.3 (2019): 283-293.

[7] Zhu, R, Zhang, T, Zhao, Q. A transfer learning approach to cross-database facial expression recognition [C]. International Conference on Biometrics.2015:293-298.

[8] Wang, F, Xiang, X, Liu, C. Transferring face verification nets to pain and expression regression. arXiv preprint arXiv, 2017.

[9] Zhai Y, Liu J, Zeng J, et al. Deep convolutional neural network for facial expression recognition [C]. International Conference on Image and Graphics, 2017:211-223.

[10] Frénay B, Verleysen M. Classification in the presence of label noise: a survey [J]. IEEE Transactions on Neural Networks & Learning Systems, 2014, 25(5):845-869.

[11] Sun, N, Chen, Z, Day, R. Facial expression recognition using digitalized facial features based on active shape model [C]. Proceedings of International Conference on Computer Science. Engineering & Applications, 2016:39-46.

[12] Liu, Y, Cao, Y, Li, Y. Facial expression recognition with PCA and LBP features extracting from active facial patches [C]. Proceedings of IEEE International Conference on Real-Time Computing and Robotics (RCAR), 2016:368-373.

[13] Zhang, B, Liu, G, Xie, G. Facial expression recognition using LBP and LPQ based on Gabor wavelet transform [C]. Proceedings of IEEE International Conference on Computer and Communications, 2017:365-369.

[14] Kumar, P, Happy, S, Routray, A. A real-time robust facial expression recognition system using HOG features [C]. Proceedings of International Conference on Computing, Analytics and Security Trends, 2017:289-293.

[15] Mohammadi, M, Fatemizadeh, E, Mahoor, M. PCA-based dictionary building for accurate facial expression recognition via sparse representation [J]. Journal of Visual Communication & Image Representation, 2014,25(5):1082-1092.

[16] Zhao, Q, Zhang, D, Lu, H. Supervised LLE in ICA space for facial expression recognition [C]. Proceedings of International Conference on Neural Networks and Brain, 2005:1970-1975.

[17] Ahmed, A. Facial expression recognition using Gabor wavelet and artificial neural networks [M]. Diss. Sudan University of Science and Technology, 2016.

[18] Zhang, Y, Zong, Y, Wang, H. Facial expression recognition based on C-means and K-nearest neighbor algorithms [J]. Caai Transactions on Intelligent Systems, 2008.

[19] Wang, X, Liu, A, Zhang, S. New facial expression recognition based on FSVM and KNN [C]. International Journal for Light and Electron Optics, 2015,126(21):3132-3134.

[20] Hai, T, Le, H, Thuy, N. Facial expression classification using artificial neural network and K-Nearest neighbor [J]. International Journal of Information Technology & Computer Science, 2015,7(3):27-32.

[21] Kung, S. Facial expression recognition using Optical Flow and 3D HMM and human action recognition using cuboid and topic models [M]. Oakland USA: Oakland University, 2016.

[22] Sumeet, S, Singh, S, Saini, R. Hardware accelerator for facial expression classification using linear SVM [J]. Advances in Signal Processing and Intelligent Recognition Systems, 2016:39-50.

[23] Liu, S, Chen, X, Fan, D. 3D smiling facial expression recognition based on SVM [C]. Proceedings of IEEE International Conference on Mechatronics and Automation, 2016:1661-1666.

[24] Liu, P, Han, S, Meng, Z. Facial expression recognition via a boosted deep belief network [C]. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2014:1805-1812.

[25] Liu, M, Li, S, Shan, S. AU-inspired deep networks for facial expression feature learning [J]. Neurocomputing, 2015,159(2):126-136.

[26] Lecun, Y, Boser, B, Denker, J. Backpropagation applied to handwritten zip code recognition [J]. Neural Computation, 2008,1(4):541-551.

[27] Burkert, P, Trier, F, Afzal, M. Dexpression: deep convolutional neural network for expression recognition. arXiv preprint arXiv, 2015.

[28] Yu, Z, Zhang, C. Image based static facial expression recognition with multiple deep network

learning [C]. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, 2015:435-442.

[29] Zhang, X. Facial expression analysis via transfer learning [M]. Diss. University of Denver, 2015.

[30] Chen, J, Liu, X, Tu, P. Person-specific expression recognition with transfer learning [C]. Proceedings of IEEE International Conference on Image Processing, 2012:2621-2624.

[31] Wang, F, Xiang, X, Liu, C. Transferring face verification nets to pain and expression regression. arXiv preprint arXiv,2017.

[32] Ding, H, Zhou, S, Chellappa, R. Facenet2expnet: regularizing a deep face recognition net for expression recognition [C]. Procedings of the 12th IEEE International Conference on Automatic Face & Gesture Recognition, 2017:118-126.

[33] Yu K, Wang Z, Zhuo L, et al. Harvesting web images for realistic facial expression recognition [C]. International Conference on Digital Image Computing: Techniques and Applications, 2010:516-521.

[34] Sukhbaatar S, Bruna J, Paluri M, et al. Training convolutional networks with noisy labels [J]. Computer Science, 2014.

[35] Wu, X, He, R, Sun, Z. A lightened CNN for deep face representation. Computer Science, 2015:111-118.

[36] Pan, S, Yang, Q. A survey on transfer learning [C]. Proceedings of IEEE Transactions on Knowledge & Data Engineering, 2010,22(10):1345-1359.

[37] Zhou Z, Shin J, Gurudu S, et al. AFT*: integrating active learning and transfer learning to reduce annotation efforts [J]. arXiv preprint arXiv, 2018.

[38] Viola P, Jones M. Robust real-time face detection [C]. Eighth IEEE International Conference on IEEE Xplore, 2004:747.

[39] Mollahosseini, A, Chan, D, Mahoor, M. Going deeper in facial expression recognition using deep neural networks [J]. Computer Science, 2015:1-10.

[40] Levi, G, Hassner, T. Emotion recognition in the wild via convolutional neural networks and mapped binary patterns [C]. Proceedings of ACM on International Conference on Multimodal Interaction, 2015:503-510.

[41] Ng, H, Nguyen, V, Vonikakis, V. Deep learning for emotion recognition on small datasets

using transfer learning [C]. Proceedings of ACM International Conference on Multimodal Interaction, 2015:443-449.