Pattern Recognition Letters 125 (2019) 105–112

Contents lists available at ScienceDirect

Pattern Recognition Letters

journal homepage: www.elsevier.com/locate/patrec

Facial expression recognition boosted by soft label with a diverse ensemble[☆]

Yanling Gan, Jingying Chen*, Luhui Xu

National Engineering Research Center for E-Learning, Central China Normal University, 152 Luoyu Road, Wuhan 430079, China

ARTICLE INFO

Article history:
Received 22 November 2018
Available online 8 April 2019

Keywords:
Facial expression recognition
Convolutional neural network
Soft label
Label-level perturbation strategy
Ensemble classifier

ABSTRACT

Facial expression recognition (FER) has recently attracted increasing attention with its growing applications in human-computer interaction and other fields. But a well-performing convolutional neural network (CNN) model learned using hard label/single-emotion label supervision may not obtain optimal performance in real-life applications because captured facial images usually exhibit expression as a mixture of multiple emotions instead of a single emotion. To address this problem, this paper presents a novel FER framework using a CNN and soft label that associates multiple emotions with each expression. In this framework, the soft label is obtained using a proposed constructor, which mainly involves two steps: (1) training a CNN model on a training set using hard label supervision; (2) fusing the latent label probability distribution predicted by the trained model to obtain soft labels. To improve the generalization performance of the ensemble classifier, we propose a novel label-level perturbation strategy to train multiple base classifiers with diversity. Experiments have been carried out on 3 publicly available databases: FER-2013, SFEW and RAF. The results indicate that our method achieves competitive or even better performance (FER-2013: 73.73%, SFEW: 55.73%, RAF: 86.31%) compared to state-of-the-art methods.

1. Introduction

Expression conveys an important visual cue to the human emotional state and intention. Six basic expressions, including anger, disgust, fear, happiness, sadness and surprise are considered as universal emotional expressions among all people [7]. Automatic facial expression recognition (FER) enables a large variety of applications in human-computer interaction and other fields [3,4], which has attracted the attention of many researchers with different backgrounds, from education to computing and psychology, etc. In developing a FER system, traditional methods usually involve 2 crucial steps: feature extraction and expression recognition. However, deep learning (DL) is more attractive because it performs these 2 steps in a unified framework and allows end-to-end training.

According to Li et al. [19], expression in images captured in real-life scenarios usually exhibits a combination/mixture of multiple basic emotions instead of a single emotion, and each emotion presents with different intensities. However, many existing DL-based FER algorithms use a single-emotion label or hard label as supervision information. These algorithms may not obtain optimal performance in real-file applications, since hard label could not characterize the correlation/ambiguity among different emotions, hence is not well suitable for describing real-life expressions. Additionally, these algorithms are prone to over-fitting in scenarios where large-scale training data are routinely unavailable. By contrast, soft label is rarely used for addressing FER problems. Soft label allows an instance to be annotated with more than one label, so it can provide more supervision information for each training sample, enabling the deep model to be trained effectively on small databases while avoiding over-fitting [13]. More appealingly, soft label can provide more intuitive descriptions for facial expression images that present multiple emotions. Given an image, soft label annotates it using a positive score vector, where each score element denotes the intensity of a basic emotion. The scores also imply the relevant/ambiguous level between different basic emotions. Fig. 1 shows some expression samples and their associated soft labels obtained by our proposed soft label constructor. As shown, the largest scores correspond to the ground-truth classes. The greater intensity that the non-ground-truth class has, the higher the correlation/ambiguity between it and the ground-truth class. Since expression usually presents a combination or mixture of the 6 basic emotions, it is beneficial to use soft label as supervision information to develop an effective FER system that performs well in the real world. This also agrees with previous researches

[☆] **Conflict of interest.** The authors declare no conflict of interest.

* Corresponding author.
E-mail address: chenjy@mail.ccnu.edu.cn (J. Chen).

https://doi.org/10.1016/j.patrec.2019.04.002
0167-8655/© 2019 Published by Elsevier B.V.

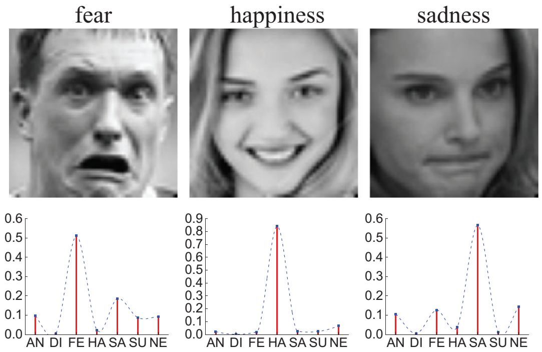

Fig. 1. Facial expression images (top) and their associated soft labels obtained by the proposed constructor (bottom).

that learning with proper ambiguity contributes to improving machine learning performance [13].

In this paper, we present a novel FER framework based on deep CNNs and soft labels. In this framework, we first train a CNN model using hard label supervision with a traditional softmax loss function, and the soft labels are subsequently generated by fusing the prediction latent label probabilities from the trained model. This whole process is referred to as soft label constructor. To increase the diversity of the base classifiers for the ensemble, we further adopt a label-level perturbation strategy, where the constructed soft labels are perturbed by a perturbation factor while keeping the relative relationships among the expressions. To the best of our knowledge, this is the first work where an ensemble using label-level perturbation is studied. In summary, the contributions of our work are threefold:

1. We propose a novel soft label construction method to construct soft labels that could describe the natural correlation/ambiguity among expressions, obtaining superior performance compared to hard labels.
2. For the diverse ensemble of the base classifiers, we proposed label-level perturbation strategy, where the constructed soft labels are perturbed by a perturbation factor to generate soft labels with different smooth degrees as supervision information, contributing to increasing the diversity of the base classifiers and, thus, improving the final FER performance by ensemble.
3. Compared to state-of-the-art methods, our method obtains competitive or even better performance on 3 benchmark FER databases: FER-2013 [12], SFEW [5] and RAF [19].

2. Related works

Facial expression recognition (FER) aims to identify the expressions in images according to their appearance/geometry features. CNNs have gained great popularity for developing FER frameworks due to their superior performance in learning features in practice. Jung et al. [14] proposed a joint fine-tuning strategy in a deep framework containing 2 different CNN networks. One extracts temporal appearance features from input video sequences, while the other extracts temporal geometry features from the trajectories of landmarks. The outputs of these 2 networks are combined using a joint fine-tuning strategy for FER. Liu et al. [22] combined (N+M)-tuplet cluster loss and softmax loss in one framework via joint optimization, and an identity-aware hard-negative mining strategy enables their method to be identity-invariant. Liu et al. [21] proposed AU-aware Deep Networks (AUDN) to extract features from face images, and then they used the extracted features to train a linear SVM classifier for expression recognition. Li et al. [19] proposed a deep locality-preserving CNN (DLP-CNN) method, where a locality-preserving loss is used to improve the discrimination ability of deep features.

Soft label allows each instance to be annotated with multiple labels. There are also many similar works where each instance is annotated with a supervision vector [9,10,29]. To avoid over-fitting and better describe the smooth transition characteristic of head poses, Geng et al. [9] used a bivariate Gaussian function to construct label distributions as supervision information and successfully applied them in head pose estimation task. Constructing label distributions as supervision also delivers promising results in many other machine learning fields [10]. Wang et al. [29] developed distributed labels by utilizing generative adversarial network (GAN) for image expressions to address the cross-database recognition problem. To alleviate the noisy label problem, Barsoum et al. [1] relabeled the FER-2013 database with the crowd-sourced label distribution according to the manual annotation results of 10 taggers, and they obtained a good performance using the new database. In general, soft label has revealed its effectiveness in many fields, but for FER tasks, it requires a lot of manual labeling [1]. In this paper, we proposed a CNN-based constructor to obtain soft labels as supervision. We use the expression training set with their hard labels to train the base architecture of the constructor and then obtain soft labels by fusing the latent label probability distribution predicted by the trained model. Soft label reveals the relevant/ambiguous level between different basic emotions and, thus, can provide much more useful information for each training sample than hard label, contributing to improving FER performance [13].

Diverse ensemble enhances a model's discrimination performance by training diverse classifiers and then combining them using some strategy, such as unweighted score/probability averaging, majority voting and so on, to make predictions. It has been studied extensively for addressing FER problems while combining CNN models [16,17,30,34]. To inject diversity for an ensemble, the existing methods usually adopt 2 perturbation strategies: input-level perturbation and network-level perturbation. The first strategy usually perturbs the inputs using different preprocessing, different features, and so on. For example, Kim et al. [16] adopted various normalization techniques including min-max normalization, isotropic diffusion-based normalization, and histogram equalization, as well as translation and rotation, to preprocess input face sets. Levi et al. [17] developed Local Binary Patterns (LBP) code mapping from the LBP features as input representations of CNN and used them to train an ensemble classifier. The second strategy usually varies the network weight initialization, network parameter, as well as the network architecture and so on, so we refer to such a strategy as network-level perturbation, and for some specific methods one can refer to [18,26]. Pons et al. [26] designed CNNs using different sizes of filters and a different number of neurons in the fully connected layer, etc. And they initialized the CNN weights with different pre-trained models. Li et al. [18] generated CNNs using random parameters and structures, including the number of layers, as well as the kernel shape size, the pooling kernel size, the number of feature maps, the learning rate, etc. However, few studies generate ensemble diversity using perturbing labels, which can be viewed as label-level perturbation. The previous works [27,32] only utilize label perturbation for either data enhancement or increasing the regularization capability of a deep model. Xie et al. [32] proposed DisturbLabel algorithm where a part of labels are randomly replaced by incorrect values in each iteration and empirically showed that DisturbLabel can be interpreted as a regularization method for preventing over-fitting. Sanchez-Lozano et al. [27] perturbed AU intensity labels by multiplying a small random number in the training process for data augmentation. Differing from the existing methods [16,17,30,34], we, for the first time, utilize label-level perturbation to increase the classifier diversity for the ensemble, where we develop a new soft label perturbation strategy.

3. Methodology

In this section, we first introduce our overall framework and then elaborate on the soft label construction method as well as the label-level perturbation strategy for the ensemble.

3.1. The overview of the proposed framework

Our framework boosts the FER performance via soft labels with a diverse ensemble. In the training stage, we first propose a constructor involving a two-step scheme to obtain soft labels: (1) train a CNN model using hard labels as supervision information and softmax loss as an optimization function, and (2) subsequently fuse the prediction probability from the trained model to obtain soft labels. Then the soft labels with different perturbation factors and Kullback–Leibler (KL) divergence loss are used to train multiple CNN base classifiers with diversity. In the inference stage, we make an ensemble of these base classifiers to make a strong prediction for expression recognition. The overall framework of the proposed method is summarized in Fig. 2. We adopt VGG16 as the base architecture for our soft label constructor and base classifier. It contains 13 convolution layers and 3 fully connected layers, where the convolutional layers are divided into 5 groups each followed by a max pooling layer and outputs feature maps with 1/32 the resolution of the input image, and the fully connected layers transform the feature maps into class scores. Our base architecture is identical to VGG16, but the last fully connected layer is modified as C-way outputs, where C is the number of the expression categories.

3.2. Soft label constructor

The constructed soft labels should be able to reflect the latent similarity or combination/mixture of relationships among expressions. For example, anger and fear, in many real-life scenarios, are usually interrelated. That is to say, they may happen simultaneously but with different intensity. In the constructor, the base model is trained using a training set with hard labels by optimizing the traditional softmax loss. Then the trained model predicts the validation set to obtain the latent label probability distribution, which, by fusing, we can obtain the soft label. The main intuition behind this is that the trained model transforms the validation set to latent C-dimensional label probability space, which implies a universal expression distribution philosophy; therefore, fusing the latent label probability distribution obtains soft labels that can describe the combination/mixture characteristics of expressions.

Formally, the constructor for soft labels is a mapping $f_{constructor} : (\mathcal{T}, \mathcal{V}) \rightarrow \mathbf{D}$, where $\mathcal{T}$ and $\mathcal{V}$ stand for the training set and validation set respectively, and $\mathbf{D} \in R^{C\times C}$ is a matrix where each row corresponds to the constructed soft labels for one class. Given the training set $\mathcal{T}$ with hard labels, we train the CNN model for the constructor using a softmax loss optimization function, and then the mapping from input image I to a C-dimensional probability outputs $\mathbf{q}$ from the trained CNN model can be denoted as:

$$\mathbf{q} = f_q(\theta, I), \tag{1}$$

where θ is the learned model parameters.

We divide the validation set $\mathcal{V}$ into C subsets according to expression category, and the jth subset for jth class expression is represented as $\mathcal{V}_j$. The unnormalized soft labels for the jth class expression can be computed by fusing the latent label distributions of all the samples from $\mathcal{V}_j$:

$$\bar{\mathbf{q}}_j = \frac{1}{|\mathcal{V}_j|} \sum_{I_i \in \mathcal{V}_j} \mathbf{q}_i \tag{2}$$

where $|\cdot|$ represents the size of the set. As with other soft label construction strategies [33], the constructed soft labels should satisfy that the sum of its element is 1; hence the normalized soft labels are:

$$\mathbf{d}_j = \frac{\bar{\mathbf{q}}_j}{||\bar{\mathbf{q}}_j||_1} \tag{3}$$

where $\mathbf{d}_j$ is the jth column in $\mathbf{D}$. Note that the maximal element in $\mathbf{d}_j$ should correspond to the ground-truth category, namely, satisfying the equality: $j = \arg\max_k(d_{j,k}), d_{j,k} \in \mathbf{d}_j$.

The elements in $\mathbf{d}_j$ describe the relationships between the jth class expression and other expressions, which should follow some

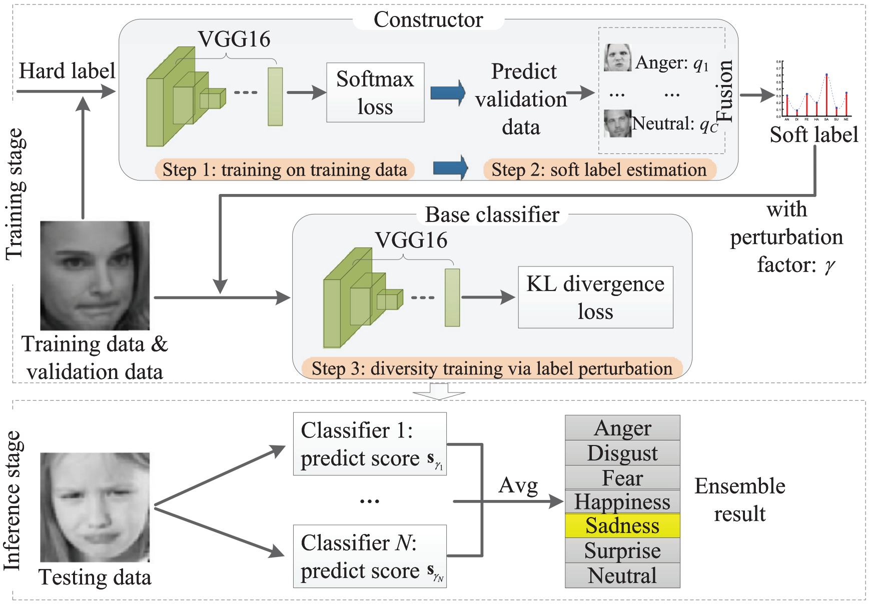
Constructor
VGG16
Hard label
Softmax loss
Predict validation data
Anger: q_1
...
Neutral: q_C
Fusion
Soft label
Step 1: training on training data
Step 2: soft label estimation
Training stage
Base classifier
VGG16
KL divergence loss
with perturbation factor: γ
Training data & validation data
Step 3: diversity training via label perturbation
Inference stage
Testing data
Classifier 1: predict score s_{γ_1}
...
Classifier N: predict score s_{γ_N}
Avg
Anger
Disgust
Fear
Happiness
Sadness
Surprise
Neutral
Ensemble result

Fig. 2. The overview of the proposed framework. In training stage, we first train a traditional softmax classifier on training data (Step 1), and then use the trained model to predict validation data to obtain the soft labels (Step 2). Finally, we train multiple classifiers with different perturbation factors (Step 3). In inference stage, the score outputs of multiple classifiers are fused to make the final decision.

rules. For example, assume the jth expression stands for fear, then the value corresponding to sadness in the vector should be greater than that corresponding to happiness. The reason for this is that sadness should be closer to fear than happiness. To keep such correlation among expressions and to control the perturbation level of the constructed soft labels, we introduce a smooth hyper-parameter γ, called perturbation factor. Then the soft labels are perturbed via the following formula:

$$\mathbf{d}_j := \frac{1}{Z}(\mathbf{d}_j)^{\gamma}, \tag{4}$$

where $(\cdot)^{\gamma}$ denotes the element-wise power operation in the vector, and Z is the normalization factor to ensure $\mathbf{d}_j$ satisfies the constraint of sum to 1. γ controls the smooth degree of the soft labels without changing the relative relationship among expressions. The smaller γ is, the smoother the label distribution is, and vice versa.

The underlying idea of the soft label is similar to that of center loss [31], which pushes the deep features of samples toward their class centers during training. Soft labels can be regarded as features since the final labels are generated by applying another argmax operation, and it also forces samples toward their corresponding soft labels that can be viewed as class centers. Thus, their underlying ideas are similar. But center loss often needs to be combined with an external classification loss, such as cross entropy, to optimize, while soft label only utilizes single loss such as KL divergence. Besides, the class centers in center loss are randomly initialized and updated iteratively. In contrast, our soft labels are generated by pre-trained model, which is more semantically meaningful.

3.3. Base classifier learning

Compared to hard label, soft label contains more useful information that describes the correlation/ambiguity between different expression categories. We use the proposed method to construct soft labels and then utilize them as supervision to train base classifiers. Therefore, the goal of base classifier training is to minimize the following KL divergence loss:

$$L = -\frac{1}{M}\sum_{i=1}^{M}\sum_{k=1}^{C} d_{y_i,k}^{(i)} \log p_k^{(i)} \tag{5}$$

where M is the training batch size, y_i is the ground-truth label of the ith sample, $d_{y_i,k}^{(i)}$ is the kth element in the soft label of the ith sample, and $p_k^{(i)}$ is the normalized logit from the base classifier and is defined as follows:

$$p_k^{(i)} = \frac{e^{s_k^{(i)}}}{\sum_{m=1}^{C} e^{s_m^{(i)}}}. \tag{6}$$

3.4. Diverse ensemble

The learning diversity of base classifiers routinely improves the generalization performance of the ensemble classifier. To inject diversity, we use a series of perturbation factors $\{\gamma_t | t = 1, \ldots, N\}$, investigated in Section 4.3, to perturb the soft labels and then train N base classifiers. To combine the decisions from these classifiers for inference, we use the unweighted score average rule, which is a simple yet effective ensemble approach. For a new test sample $\mathbf{x}$, assume that the tth base classifier outputs a C-dimensional score vector, denoted by $\mathbf{s}_{\gamma_t}$, through the last full connected layer. We directly average over $\mathbf{s}_{\gamma_t}$, and the class with the highest average score is reported as the predicted class. So, the prediction label y for $\mathbf{x}$ can be formulated as follows:

$$y = \arg\max_j \left(\frac{1}{N}\sum_t \mathbf{s}_{\gamma_t}\right). \tag{7}$$

4. Experiments

In this section, we first investigate the label perturbation factor γ for training an ensemble classifier and then report the experimental results for the proposed method, as well as a comparison with state-of-the-art methods.

4.1. Datasets

We evaluate the proposed method on FER-2013 [12], SFEW [5] and RAF [19], which are widely used as benchmark databases for FER. Some samples are shown in Fig. 3.

FER-2013 was released for FER competition in ICML 2013 challenges in representation learning [12]. This database contains real-life images with diverse backgrounds, illumination conditions, etc. There are a total of 35,887 face images with 7 expressions (6 base expressions and 1 neutral expression): anger (AN), disgust (DI), fear (FE), happiness (HA), sadness (SA), surprise (SU), and neutral (NE). The faces in this database have been cropped and automatically registered. We resize them to 224 × 224 pixels as inputs for our model. The competition organizers divide the database into 3 parts: a training set with 28,709 images, a public test set (validation set) with 3589 images, and a private test set (test set) with 3589 images. We use the training set and public test set for training and report recognition accuracy on the private test set.

SFEW was introduced for a sub-challenge in EmotiW 2015 [5]. There are a total of 1766 images in this database and they are split into 3 parts, i.e., 958 for training, 436 for validation, and 372 for testing. Each image has been annotated with one of the 7 expression categories. In this database, face regions are detected and registered with respect to 5 landmark positions using a face detector in [35]. Then they are cropped and normalized to 224 × 224 pixels. Because the annotations of the test set are not publicly available, we train on the training set and report the recognition accuracy on the validation set, as many methods did [16,17,19,24,26].

RAF is a large-scale expression database collected from the Internet, including two subsets: single-label subset and two-tab subset [19]. The first subset is used in our experiment, and it contains 15,339 images annotated with 7 expression categories and has been split into two groups: 12,271 images for training and 3068 images for testing. We use the automatically detected landmarks to align images and crop the face regions of the size of 224 × 224.

4.2. Implementation detail

Our method is implemented using Caffe on the GTX 1080 Ti. Our base architecture for the constructor and the base classifier are initialized by VGG-Face [25]. The dropout ratio, momentum, weight decay, and batch size are set to 0.5, 0.9, 0.0005, and 48, respectively. On the SFEW database, the total training number is set to 0.5K, and the learning rate is fixed at 0.01. On the FER-2013

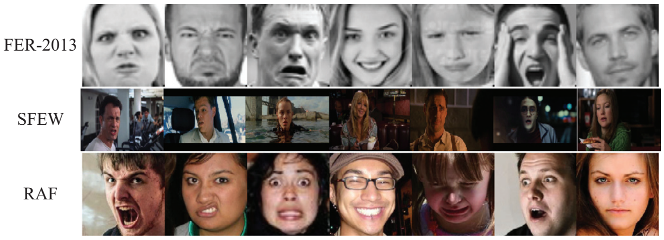

Fig. 3. Examples from the FER-2013, SFEW and RAF databases. The images from left to right correspond to anger, disgust, fear, happiness, sadness, surprise, and neutral respectively.

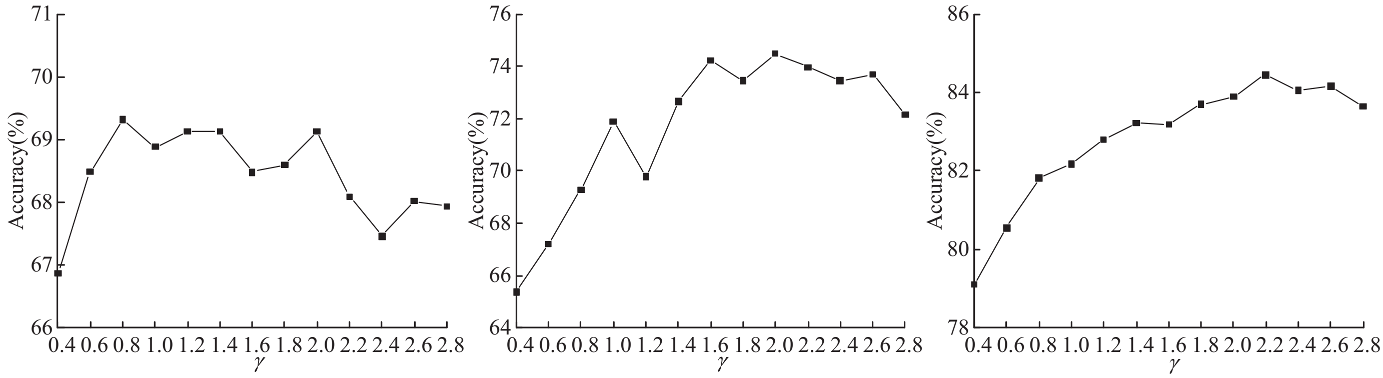

Fig. 4. Investigation results on FER-2013 (left), SFEW (middle) and RAF (right).

Table 1
The constructed soft labels for each category.

AN	0.6247	0.0136	0.1023	0.0419	0.1069	0.0202	0.0905
DI	0.2174	0.5696	0.0833	0.0090	0.1147	0.0001	0.0059
FE	0.0977	0.0038	0.5115	0.0224	0.1858	0.0869	0.0919
HA	0.0226	0.0009	0.0193	0.8416	0.0245	0.0235	0.0675
SA	0.1060	0.0051	0.1269	0.0389	0.5671	0.0123	0.1437
SU	0.0269	0.0004	0.0870	0.0464	0.0163	0.8000	0.0230
NE	0.0655	0.0001	0.0740	0.0769	0.1673	0.0133	0.6030

Table 2
Comparison of different methods on FER-2013.

Method	Accuracy (%)	
	Ensemble	Single classifier
Our method (γ perturbation)	73.73	71.47
Our method (γ = 1.4)	73.12	
Our method (γ = 1.0)	72.95	71.02
VGG-Face [25]	72.67	69.10
Kim et al. [16]	72.72	70.58
Winning method in [12]	–	71.16
Wen et al. [30]	69.96	–
Ng et al. [24]	–	51.1
Li et al. [18]	70.66	67.15

database, we start with a learning rate of 0.01 and decrease it by 0.1 after each 5K iterations. The total training number is set to 8K. On the RAF database, we start with a learning rate of 0.01 and decrease it by 0.1 after each 3K iterations. The maximum iteration is 6K. We use the FER-2013 database to construct the soft labels with the proposed constructor as described in Section 3.2. Specifically, we use the training set to fine-tune the pre-trained base model for the constructor, which achieves accuracies of 67.46% and 69.38% on the validation set and test set respectively, then we construct the soft labels for each category according to the predicted results on the validation set, and the constructed soft labels are shown in Table 1. Then the constructed soft labels are used as supervision for expression recognition experiments on the FER-2013, SFEW and RAF databases.

4.3. Parameter investigation

In general, injecting diversity can make the ensemble learning more capable of enhancing the model's discrimination ability. For this purpose, we perturb the soft labels with different values of perturbation factor γ to train the base classifiers. We first investigate the effect of γ with a series of experiments by assigning $\gamma \in$ [0.4, 0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0, 2.2, 2.4, 2.6, 2.8]. For FER-2013 database, we use the training set for base classifier training and the public test set for validation. Because SFEW is a small-scale database, we enlarge its training set by horizontal mirroring and then perform the investigation on the enlarged training set, of which 80% are selected for training, and the rest are selected for validation. For RAF database, we select 80% of the training set for training, and the remaining 20% for validation. The investigation results are shown in Fig. 4. On the FER-2013 database, the accuracy is relatively low at point $\gamma = 0.4$. One possible reason is that the small parameter value makes soft labels too smooth to distinguish different classes. When γ is in the range of 0.6–2.0, the base classifiers obtain a better validation accuracy above 68.4%. After the point $\gamma = 2.0$, the accuracy starts to drop dramatically. Therefore, on the FER-2013 database, we select $\gamma \in$ [0.6, 0.8, 1.0, 1.2, 1.4, 1.6, 1.8, 2.0] to train $N = 8$ base classifiers. On the SFEW database, in general, the accuracy increases as γ increases, and it presents a better performance when γ is in the range of 1.6–2.6. So, we select $\gamma \in$ [1.6, 1.8, 2.0, 2.2, 2.4, 2.6] to train $N = 6$ different base classifiers for the ensemble. Similarly, the γ in the range of [1.8, 2.0, 2.2, 2.4, 2.6, 2.8] achieves relatively superior performance on RAF database, so we select this range for a final ensemble.

4.4. Comparison with other methods on the FER-2013 database

Our method utilizes VGG-Face to initialize the base classifiers, so we fine-tune it with softmax loss as a baseline. Besides, we also compare the proposed method with state-of-the-art methods. Tang used an L2-SVM loss function to train a deep FER model, and they won the first place in FER challenge with an accuracy of 71.16% [12]. Ng et al. [24] fine-tuned existing advanced architectures on FER-2013, which are pre-trained on a large-scale external database, finally showing an accuracy of 51.1%. The methods in [16,18,30] showed that ensemble is effective in improving the generalization performance of models. Kim et al. [16] proposed a hierarchical committee for assembling multiple CNN classifiers and finally obtained a best single model accuracy of 70.58% and an ensemble accuracy of 72.72%. Li et al. [18] proposed maximum relevance and minimum redundancy-based ensemble pruning to select the most capable base classifiers while minimizing the redundancy among these classifiers. They obtained a best single classifier accuracy of 67.15% and an ensemble accuracy of 70.66%. Wen et al. [30] presented a probability-based fusing method for a CNN classifier ensemble.

We train 8 base classifiers using diverse smoothed soft labels that are investigated in Section 4.3 and make an ensemble of these classifiers using the average rule for expression recognition. Our single classifier (i.e., $\gamma = 1.4$) achieves an accuracy of 71.47% that outperforms the implemented base CNN (VGG-Face) and other single classifiers [12,16,18,24,25]. Further, our diversity ensemble achieves 73.73% that outperforms the ensemble results of 8 classifiers trained without factor perturbation (i.e., $\gamma = 1.4$ or 1.0) and

Table 3
Confusion matrix on FER-2013 (%).

Predicted label								
		AN	DI	FE	HA	SA	SU	NE
True label	AN	65.38	0.20	9.57	1.83	16.09	0.41	6.52
	DI	27.27	63.64	5.45	1.82	1.82	0.00	0.00
	FE	10.04	0.00	53.41	2.46	22.54	6.25	5.30
	HA	1.25	0.00	1.25	90.33	1.82	1.59	3.75
	SA	6.73	0.00	7.41	3.20	69.36	0.17	13.13
	SU	1.92	0.00	7.69	4.81	1.68	82.45	1.44
	NE	3.04	0.00	2.08	3.19	17.57	0.80	73.32

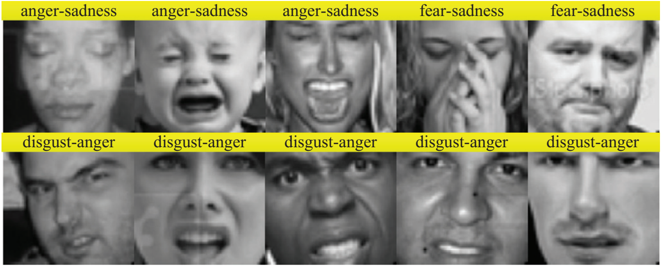

Fig. 5. Examples of false prediction and their associated labels (true labels-predicted labels) on the FER-2013 database.

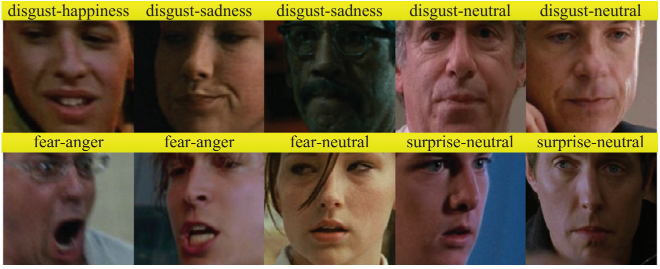

Fig. 6. Examples of false prediction and their associated labels (true labels-predicted labels) on the SFEW database.

Table 4
Comparison of different methods on SFEW.

Method	Accuracy (%)	
	Ensemble	Single classifier
Our method (γ perturbation)	55.73	52.75
Our method ($\gamma = 2.2$)	53.21	
Our method ($\gamma = 1.0$)	52.98	51.61
VGG-Face [25]	48.62	43.81
Kim et al. [16]	56.4	52.5
Levi et al. [17]	51.75	44.73
Li et al. [19]	–	51.05
Ng et al. [24]	–	48.5
Pons et al. [26]	42.9	41.3
Cai et al. [2]	–	52.52
Meng et al. [23]	–	50.98
Kaya et al. [15]	53.06	–
Ding et al. [6]	–	55.15
Sun et al. [28]	48.94	–

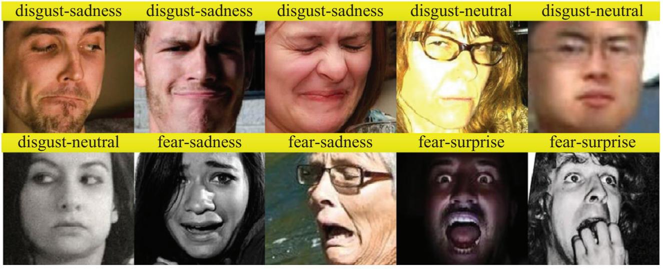

Fig. 7. Examples of false prediction and their associated labels (true labels-predicted labels) on the RAF database.

Table 5
Confusion matrix on SFEW (%).

Predicted label								
		AN	DI	FE	HA	SA	SU	NE
True label	AN	72.73	3.90	2.60	7.79	1.30	1.30	10.39
	DI	8.70	13.04	8.70	13.04	17.39	8.70	30.43
	FE	21.28	0.00	31.91	4.26	10.64	4.26	27.66
	HA	2.74	0.00	0.00	87.67	4.11	0.00	5.48
	SA	2.74	0.00	9.59	9.59	52.05	5.48	20.55
	SU	29.82	0.00	10.53	14.04	3.51	19.30	22.81
	NE	9.30	0.00	3.49	3.49	13.95	4.65	65.12

the state-of-the-art ensemble classifiers [16,18,30], confirming the effectiveness of our ensemble strategy.

To get a better understanding of the proposed method, Table 3 shows the confusion matrix, and Fig. 5 shows some false prediction examples. It appears that the accuracies of happiness and surprise are relatively high. Meanwhile, disgust is easily misclassified as anger, while anger and fear tend to be misclassified as sadness, which may be caused by the fact that they are often accompanied by sadness.

4.5. Comparison with other methods on the SFEW database

Table 4 shows the performance comparison between our method and state-of-the-art methods. Ng et al.[24] and Ding et al. [6] adopted transfer learning technology to address the FER problem, obtaining accuracies of 48.50% and 55.15%, respectively. But these methods require large-scale external expression database. The methods in [2,19,23] made effort to design loss function to learn discriminative deep expression features. Many methods [15–17,26,28] assembled multiple classifiers for boosting FER performance. Kaya et al. [15] and Sun et al. [28] extracted multiple types of features as classifier inputs for ensemble. Pons et al. [26] proposed a supervised committee for a trainable ensemble of CNN classifiers (64 VGG16 or 72 4-layer traditional CNNs), but they obtained poor results. One possible reason for such results is that assembling excessive amounts of classifiers may lead to performance degradation. Levi et al. [17] developed mapped LBP as CNN inputs, showing a best single classifier accuracy of 44.73% and an ensemble accuracy of 51.75%. But it can be seen in Table 4, this method provided inferior results to ours. Kim et al. [16] hierarchically combined 216 deep classifiers and obtained an accuracy of 56.4%, which is higher than ours. But their best single model performance is inferior to ours. Besides, to handle the small size problem in this database, they used 2 external facial expression databases along with the SFEW data for classifier training.

Table 6
Comparison of different methods on RAF.

Method	Accuracy (%)	
	Ensemble	Single classifier
Our method (γ perturbation)	86.31	85.20
Our method ($\gamma = 2.2$)	85.82	
Our method ($\gamma = 1.0$)	84.97	84.55
VGG-Face [25]	85.43	83.25
Li et al. [20]	–	83.27
Li et al. [19]	–	74.20
Fan et al. [8]	76.73	–
Ghosh et al. [11]	–	77.48

Table 7
Confusion matrix on RAF (%).

Predicted label								
		AN	DI	FE	HA	SA	SU	NE
True label	AN	77.16	1.85	0.00	8.02	2.47	3.09	7.41
	DI	10.63	44.38	0.00	9.38	13.13	2.50	20.00
	FE	4.05	0.00	51.35	8.11	12.16	20.27	4.05
	HA	0.17	0.17	0.00	95.36	0.59	0.51	3.21
	SA	0.63	0.84	0.21	5.65	82.01	0.00	10.67
	SU	1.52	0.91	0.61	2.74	1.22	86.32	6.69
	NE	0.29	0.44	0.00	5.29	3.53	1.03	89.41

On this database, we make an ensemble of 6 classifiers trained using smoothed soft labels, as mentioned in Section 4.3. Our single classifiers (i.e., $\gamma = 2.2$ and $\gamma = 1.0$) obtain significant gains with respect to the VGG-Face classifier and outperforms most of the state-of-the-art single classifiers, showing the advantages of soft label. Moreover, our method achieves an ensemble accuracy of 55.73%, showing better performance compared to the ensemble result of 6 classifiers trained using fixed perturbation factor $\gamma = 2.2$ and other ensemble methods. This also illustrates that our method is beneficial to improving FER performance.

The confusion matrix of our method on the validation set of SFEW is reported in Table 5. In this database, happiness and anger are easily distinguishable. Meanwhile, fear, surprise, and disgust tend to be confused with other expressions. Fig. 6 shows some examples for such false prediction cases. This phenomenon may be caused by the small number of examples available for the class and the mixing nature of expressions.

4.6. Comparison with other methods on the RAF database

The performance comparison on RAF database is given in Table 6. One can see that the ensemble performance has slight improvement without factor perturbation. By contrast, the proposed diversity ensemble method improves the performance significantly, and it outperforms the multi-region ensemble CNN [8]. Furthermore, our individual model's accuracy is higher than other existing methods [11,19,20]. Li et al. [19] released RAF database and reported the baseline accuracy as 74.20%. Ghosh et al. [11] designed a capsule network structure, and their results is inferior to the existing advanced architecture, i.e., VGG16, as shown in Table 6. Li et al. [20] used additional facial landmark information and introduced local attention mechanism, and they obtained accuracy of 83.27% on this database.

Table 7 shows the confusion matrix for the proposed method. Similar to the first two experiments (in Tables 3 and 5), happiness has the highest accuracy. Followed are neutral, surprise and sadness, which all obtain an accuracy of more than 80%. Fear and disgust show relatively lower accuracies. Disgust is prone to be misclassified as neutral or sadness and fear is prone to be misclassified as surprise or sadness, as shown in Fig. 7.

5. Conclusions

This paper proposes a novel FER framework via CNN with soft labels that associate multiple emotions to each expression image. In this framework, we first learn a soft label constructor, in which a base CNN model is trained using hard label supervision and, subsequently, soft labels are constructed by fusing the prediction outputs from the trained model. Then, a soft label perturbation strategy is developed to train diverse base classifiers to enhance the discrimination ability of the ensemble classifier. For performance evaluation, extensive experiments are conducted on the FER-2013, SFEW and RAF databases, and the proposed method achieves accuracies of 73.73%, 55.73% and 86.31% respectively. The results show that the proposed method is competitive or even better compared to the state-of-the-art methods.

Future work will focus on exploring an attention model that can emphasize the most informative regions of a face to further improve FER performance.

Acknowledgments

This work was supported by the National Key Research and Development Program of China (Grant no. 2018YFB1004504), the National Natural Science Foundation of China (Grant nos. 61772380, 61273063), Research Funds of CCNU from the Colleges' Basic Research and Operation of MOE (Grant nos. 2018CXZZ016, CCNU19Z02002). Foundation for Innovative Research Groups of Hubei Province (Grant no. 2017CFA007).

References

[1] E. Barsoum, C. Zhang, C.C. Ferrer, Z. Zhang, Training deep networks for facial expression recognition with crowd-sourced label distribution, in: Proceedings of the 18th ACM International Conference on Multimodal Interaction, ACM, 2016, pp. 279–283.

[2] J. Cai, Z. Meng, A.S. Khan, Z. Li, J. O'Reilly, Y. Tong, Island loss for learning discriminative features in facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on, IEEE, 2018, pp. 302–309.

[3] J. Chen, D. Chen, X. Li, K. Zhang, Towards improving social communication skills with multimodal sensory information, IEEE Trans. Ind. Inform. 10 (1) (2014) 323–330.

[4] J. Chen, N. Luo, Y. Liu, L. Liu, K. Zhang, J. Kolodziej, A hybrid intelligence-aided approach to affect-sensitive e-learning, Computing 98 (1-2) (2016) 215–233.

[5] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, T. Gedeon, Video and image based emotion recognition challenges in the wild: Emotiw 2015, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 423–426.

[6] H. Ding, S.K. Zhou, R. Chellappa, Facenet2expnet: regularizing a deep face recognition net for expression recognition, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 118–126.

[7] P. Ekman, W.V. Friesen, Constants across cultures in the face and emotion, J. Personal. Soc. Psychol. 17 (2) (1971) 124.

[8] Y. Fan, J.C. Lam, V.O. Li, Multi-region ensemble convolutional neural network for facial expression recognition, in: International Conference on Artificial Neural Networks, Springer, 2018, pp. 84–94.

[9] X. Geng, Y. Xia, Head pose estimation based on multivariate label distribution, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 1837–1842.

[10] X. Geng, C. Yin, Z.-H. Zhou, Facial age estimation by learning from label distributions, IEEE Trans. Pattern Anal. Mach. Intell. 35 (10) (2013) 2401–2412.

[11] S. Ghosh, A. Dhall, N. Sebe, Automatic group affect analysis in images via visual attribute and feature networks, in: 2018 25th IEEE International Conference on Image Processing (ICIP), IEEE, 2018, pp. 1967–1971.

[12] I.J. Goodfellow, D. Erhan, P.L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, et al., Challenges in representation learning: a report on three machine learning contests, in: International Conference on Neural Information Processing, Springer, 2013, pp. 117–124.

[13] G. Hinton, O. Vinyals, J. Dean, Distilling the knowledge in a neural network, (2015) arXiv:1503.02531.

[14] H. Jung, S. Lee, J. Yim, S. Park, J. Kim, Joint fine-tuning in deep neural networks for facial expression recognition, in: Proceedings of the IEEE International Conference on Computer Vision, 2015, pp. 2983–2991.

[15] H. Kaya, F. Gürpinar, S. Afshar, A.A. Salah, Contrasting and combining least squares based learners for emotion recognition in the wild, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 459–466.

[16] B.K. Kim, J. Roh, S.Y. Dong, S.Y. Lee, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimodal User Interfaces 10 (2) (2016) 173–189.

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

[18] D. Li, G. Wen, Mrmr-based ensemble pruning for facial expression recognition, Multimed. Tools Appl. 77 (12) (2018) 15251–15272.

[19] S. Li, W. Deng, J. Du, Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild, in: Computer Vision and Pattern Recognition (CVPR), 2017 IEEE Conference on, IEEE, 2017, pp. 2584–2593.

[20] Y. Li, J. Zeng, S. Shan, X. Chen, Patch-gated cnn for occlusion-aware facial expression recognition, in: 2018 24th International Conference on Pattern Recognition (ICPR), IEEE, 2018, pp. 2209–2214.

[21] M. Liu, S. Li, S. Shan, X. Chen, Au-aware deep networks for facial expression recognition., in: FG, 2013, pp. 1–6.

[22] X. Liu, B.V. Kumar, J. You, P. Jia, Adaptive deep metric learning for identity-aware facial expression recognition., in: CVPR Workshops, 2017, pp. 522–531.

[23] Z. Meng, P. Liu, J. Cai, S. Han, Y. Tong, Identity-aware convolutional neural network for facial expression recognition, in: Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on, IEEE, 2017, pp. 558–565.

[24] H.W. Ng, V.D. Nguyen, V. Vonikakis, S. Winkler, Deep learning for emotion recognition on small datasets using transfer learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 443–449.

[25] O.M. Parkhi, A. Vedaldi, A. Zisserman, et al., Deep face recognition, in: British Machine Vision Conference, 1, 2015, p. 6.

[26] G. Pons, D. Masip, Supervised committee of convolutional neural networks in automated facial expression analysis, IEEE Trans. Affect. Comput. 9 (3) (2018) 343–350.

[27] E. Sánchez-Lozano, G. Tzimiropoulos, M. Valstar, Joint action unit localisation and intensity estimation through heatmap regression, in: British Machine Vision Conference, 2018, p. 233.

[28] B. Sun, L. Li, G. Zhou, X. Wu, J. He, L. Yu, D. Li, Q. Wei, Combining multimodal features within a fusion network for emotion recognition in the wild, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 497–502.

[29] X. Wang, X. Wang, Y. Ni, Unsupervised domain adaptation for facial expression recognition using generative adversarial networks, Comput. Intell. Neurosci. 2018 (2018).

[30] G. Wen, Z. Hou, H. Li, D. Li, L. Jiang, E. Xun, Ensemble of deep neural networks with probability-based fusion for facial expression recognition, Cognit. Comput. 9 (5) (2017) 597–610.

[31] Y. Wen, K. Zhang, Z. Li, Y. Qiao, A discriminative feature learning approach for deep face recognition, in: European Conference on Computer Vision, Springer, 2016, pp. 499–515.

[32] L. Xie, J. Wang, Z. Wei, M. Wang, Q. Tian, Disturblabel: regularizing cnn on the loss layer, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016, pp. 4753–4762.

[33] L. Xu, J. Chen, Y. Gan, Head pose estimation with soft labels using regularized convolutional neural network, Neurocomputing 337 (2019) 339–353.

[34] Z. Yu, C. Zhang, Image based static facial expression recognition with multiple deep network learning, in: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, ACM, 2015, pp. 435–442.

[35] K. Zhang, Z. Zhang, Z. Li, Y. Qiao, Joint face detection and alignment using multitask cascaded convolutional networks, IEEE Signal Process. Lett. 23 (10) (2016) 1499–1503.