Information Sciences 428 (2018) 49–61

Contents lists available at ScienceDirect

Information Sciences

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

Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction

Luefeng Chen[a,b], Mengtian Zhou[a,b], Wanjuan Su[a,b], Min Wu[a,b,*], Jinhua She[a,b,c], Kaoru Hirota[d]

[a] School of Automation, China University of Geosciences, Wuhan 430074, China
[b] Hubei Key Laboratory of Advanced Control and Intelligent Automation for Complex Systems, Wuhan 430074, China
[c] School of Engineering, Tokyo University of Technology, Tokyo 192-0982, Japan
[d] Tokyo Institute of Technology, Yokohama 226-8502, Japan

ARTICLE INFO

Article history:
Received 17 April 2017
Revised 22 October 2017
Accepted 23 October 2017
Available online 1 November 2017

Keywords:
Facial emotion recognition
Deep sparse autoencoder network
Softmax regression
Human-robot interaction

ABSTRACT

Deep neural network (DNN) has been used as a learning model for modeling the hierarchical architecture of human brain. However, DNN suffers from problems of learning efficiency and computational complexity. To address these problems, deep sparse autoencoder network (DSAN) is used for learning facial features, which considers the sparsity of hidden units for learning high-level structures. Meanwhile, Softmax regression (SR) is used to classify expression feature. In this paper, Softmax regression-based deep sparse autoencoder network (SRDSAN) is proposed to recognize facial emotion in human-robot interaction. It aims to handle large data in the output of deep learning by using SR, moreover, to overcome local extrema and gradient diffusion problems in the training process, the overall network weights are fine-tuned to reach the global optimum, which makes the entire depth of the neural network more robust, thereby enhancing the performance of facial emotion recognition. Results show that the average recognition accuracy of SRDSAN is higher than that of the SR and the convolutional neural network. The preliminarily application experiments are performed in the developing emotional social robot system (ESRS) with two mobile robots, where emotional social robot is able to recognize emotions such as happiness and angry.

1. Introduction

Facial emotion recognition is the most important way of human emotion expression. For a couple of past decades, it has been a very important research field in the field of computer vision and image recognition. Nevertheless, facial expression recognition is still a challenging task [20,40]. This is mainly associated with varying lighting, pose and environment. In recent years, deep neural network (DNN) [42,27] has attracted an increasing attention in machine learning and artificial intelligence, and many types of DNN related algorithms have been successfully applied to image recognition tasks [32,24]. Being different from a shallow learning architecture [22] for single layer nonlinear transformation, DNN algorithms attempt to learn high-level abstract features in data by utilizing hierarchical architectures, which have become an effective approach

* Corresponding author.
E-mail addresses: chenluefeng@cug.edu.cn (L. Chen), wumin@cug.edu.cn (M. Wu).

https://doi.org/10.1016/j.ins.2017.10.044
0020-0255/

for extracting high-level features from data. However, traditional DNN suffers from problems of learning efficiency and computational complexity. However, the autoencoder is able to reconstruct data so that data could be better represented which can improve the efficiency of data learning. In addition, the sparsity is applied to the autoencoder and this can reduce the computational complexity of the algorithm. As a result, deep sparse autoencoder network (DSAN) is used for learning facial emotion features, and the sparsity of hidden units is utilized to support learning high-level structures. Meanwhile, Softmax regression (SR) is used to classify expression feature.

In this paper, SR-based deep sparse autoencoder network (SRDSAN) is proposed to recognize facial expression. Firstly, the regions of interest (ROI) include eyebrows, eyes, and mouth, are selected as extracted areas of facial expression images' feature. Then, greedy pre-trained network layer by layer produces initial weights, and expands the 'code' and 'decode' network, furthermore, optimizes the sparse parameters, the hidden layer nodes and the numbers of hidden layers to determine the best topology of the network. The algorithm uses the layered approach for training data and extracts different levels of data features that can build signal mapping feature from the bottom to the top. Top-level network uses SR to classify expression feature, while the gradient descent method (GD) is used to find the optimal parameters. Finally, the weights of the entire DSAN are fine-tuned via the back-propagation (BP) algorithm.

In the proposed SRDSAN, sparse representation and DNN is fused for robust facial emotion recognition, and by introducing sparsity, the feature extractor is able to learn high-level structures. Moreover, SR is used for facial emotion classification to handle the large nonlinear structure of facial images in the proposed method. Furthermore, BP algorithm is used for fine-tuning the weights of the whole SRDSAN, which can not only make the whole deep learning network more robust and enhance facial emotion recognition performance, but also can make the learning rate faster and overcome local extrema and gradient diffusion problems.

To demonstrate the effectiveness of the proposed approach, experiments are developed in an emotion recognition system, and we conducted experiments on JAFFE database and Extended CohnKanade (CK+) database. The experimental results demonstrate that the proposed SRDSAN produces the highest average accuracy of emotion recognition. Moreover, preliminary application experiments are being carried out in the developing human-robot interaction system (HRI) called emotional social robot system (ESRS). By using the ESRS, the proposal is being extended to mobile robots for analyzing and understanding human emotions, as well as responding appropriate social behaviors.

The article is structured as follows. Section 2 presents the related work of deep neural network algorithms. In Section 3, the structure and design of SRDSAN is introduced. The experimental process, some experimental results and their analysis are presented in Section 4. In Section 5, conclusions are covered.

2. Literature review

HRI [1,25] has attracted a growing interests of researchers in recent years. It is generally desired that robots could have the ability to recognize and understand human emotions. Meanwhile, the intelligent service system with emotional recognition ability becomes a hot topic in HRI. Facial expression recognition plays an important role in manifestations of recognizing and understanding human emotion by robots [14,16,23]. Traditional facial feature extraction algorithms such as Gabor wavelet transform [29], model method [2], and optical flow method [11], which are subject to a number of constraints just like face posture diversity and changeability, individual differences in facial structure and the levels of skin color, computer performance impose restrictions on the training speed, the impact on the external environment, e.g., light, shelter, and so on. Yi et al. [33] proposed a facial recognition expression algorithm by exploiting the structural characteristics and the texture information (SCTI) hidden in the image space which first marked the feature points by using active appearance model, then three facial features are proposed to eliminate the differences among the individuals, and a radial basis function neural network is utilised as the classifier. However, DNN technology [12] began to sweep the worldwide field of HRI when the traditional way of feature extraction and recognition appeared to show limitations.

Recently, the DNN has been successfully applied to the AlfaGo intelligence program which shows that the DNN has a strong ability of self-learning, and it has become a hot research topic. According to the model, deep learning and training methods can be divided into convolutional neural network (CNN) [28,18], the deep belief networks (DBN) [43], and canonical correlation analysis network [38]. Furthermore, the DNN also has shown its promise in image recognition [19,26]. Accordingly, many researchers use DNN to identify facial expressions. For example, Xie et al. [34] proposed a feature redundancy-reduced convolutional neural network (FRR-CNN) to achieve facial expression recognition, and the convolutional kernels of FRR-CNN causes divergence by presenting a more discriminative mutual difference among feature maps of the same layer, which results in generating less redundant features and yields a more compact representation of the image. In [30], local directional position pattern is applied to feature extraction process, and the proposed features are finally applied with DBN for expression recognition, the recognition performance was superior over traditional approaches. Kim et al. [15] trained multiple deep CNNs as committee members and combine their decisions for robust facial expression recognition, on public facial expression recognition databases, their hierarchical committee of deep CNNs yields superior performance. Hence, it is evident that DNN is an effective approach for facial expression recognition. For this reason, we use deep autoencoder network to realize facial emotion recognition which is also a kind of DNN.

Moreover, a special kind of face recognition method, i.e., sparse representation method (SRM) is proposed. This method supports the key idea that samples can be sparsely represented by a large number of "atoms", and it is widely used for face recognition [8,36,37]. Many ways to restrict the hidden units to be sparse have been proposed [10]. In [10], multiobjective optimization theory was used, which reconstructed error and the sparsity of hidden units, thus learning efficient features for hierarchical neural networks. A joint sparse representation and pattern learning model for robust face recognition is proposed [39], which is an efficient algorithm for face recognition. By introducing sparsity, the feature extractor is able to learn high-level structures. Therefore, we add sparsity to the deep autoencoder network.

Regression techniques, such as ridge regression (RR) [9,31] and logistic regression (LR) [13,21], have been widely used in supervised learning for pattern classification. In the recent years, RR is generalized for face recognition [3,35]. In visual classification tasks such as face recognition, the appearance of the training sample images also conveys important discriminative information. In [35], RR uses regular simplex vertices to represent the multiple target class labels, which generalizes RR to multivariate labels in order to apply it for face recognition. However, these methods mainly exploit class label information for linear mapping function learning, and they will become less effective when the number of training samples' per class is large. As a result, in this paper, SR will be used to sort the features learned from the DSAN, that is what we called SRDSAN.

However, the traditional neural network algorithm is always prone to cause local maximum and gradient diffusion problems in the training process [41], which leads to poor recognition result. Consequently, we apply greedy pre-train network layer by layer to get initial weights. Meantime, BP algorithm is used for fine-tuning the weights of the whole SRDSAN to achieve the global optimum. Accordingly, the proposal can overcome the local extrema and gradient diffusion problems when recognizing facial expression.

3. Softmax regression based deep sparse autoencoder network and functional process

3.1. Softmax regression based deep sparse autoencoder network

The structure of SRDSAN is shown in Fig. 1. It uses sparse autoencoder network for deep learning, and SR to classify expression feature.

Algorithm 1 outlines the SRDSAN method. First, eyebrows, eyes, and mouth are selected as the ROI and extracted as

Algorithm 1 SRDSAN.

1. The ROI areas clipping, and normalize the image.
2. Greedy pre-train network layer by layer to get initial weights matrix.
3. Obtain the network output.
4. Training SR to estimate the parameters.
5. Minimize cost function.
6. Fine-tune the weights of the entire SRDSAN.
7. Obtain the facial emotion.

feature of facial expression images. Then, initial weights of the network is produced by greedy pre-training the network layer by layer. To determine the best network model, hidden layer nodes and the numbers of hidden layers is obtained by optimizing the sparse parameter. Furthermore, SR is used to classify facial expression features, and the optimal model parameters of SR is trained by GD method. Finally, the weights of the entire DSAN are fine-tuned via BP algorithm, to make the whole deep learning network more robust and enhance facial emotion recognition performance.

3.2. ROI based face image preprocessing

Before realizing feature extraction, we need to complete some pretreatment processes, such as the ROI region's segmentation of facial expression images. In the facial expression images, the texture and shape of the feature changes in three key parts-eyebrows, eyes and mouth may reflect changes in facial expression. As a consequence, we can use these ROI as areas of facial expression images' feature.

For facial expression images of JAFFE database [44], we manually obtain the coordinates of the four corners in three ROI areas, and split eyebrows, eyes and mouth from these primitive facial expression images, shown in Table. 1. Information about the coordinates of four corners in three ROI areas and rectangular clipping area are listed, each column in the four corners matrix coordinates represent four points x, y coordinates shown in clockwise direction. The rectangular clipping region matrix represent the width and height of this rectangular clipping area.

By clipping these ROI areas can not only reduce the interference in facial information caused by image interference in noncritical part, but also reduce the amount of data and improve the compute speed. Specific ROI crop areas image is shown in Fig. 2.

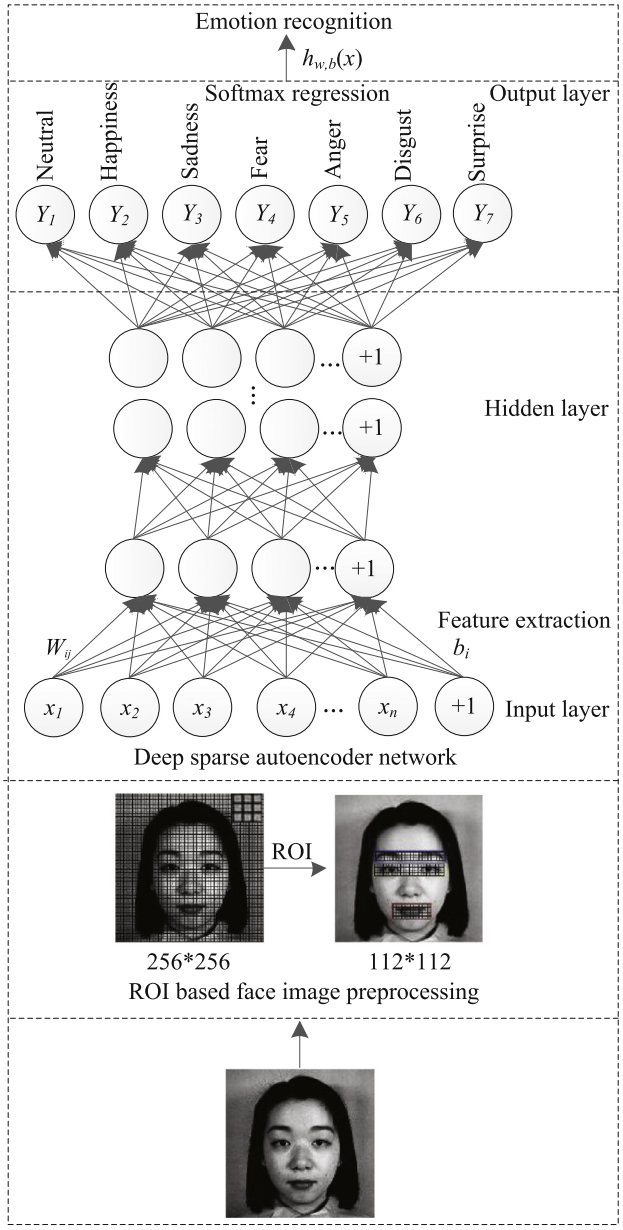

Fig. 1. Structure of SRDSAN.

Table 1
Comparison of emotion recognition experimental results

Key parts	Corners coordinates				Clipping region
	Xleftup *Xleftup*	*Xrightup* *Xrightup*	*Xleftdown* *Xleftdown*	*Xrightdown* *Xrightdown*	[*Length, Width*]
Eyebrows	74.21 100.24	182.26 100.24	182.26 120.05	74.21 120.05	[108.05 19.91]
Eyes	74.22 120.85	182.25 120.85	182.25 140.12	74.22 140.12	[108.03 19.27]
Mouth	95.55 180.03	162.01 180.03	162.01 210.05	95.55 210.05	[66.46 30.02]

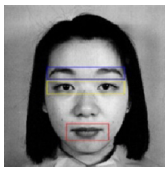

Fig. 2. ROI segment region image.

3.2.1. Expand the encode and decode network

Define v as an input layer data, h as hidden layer data. We use the trained parameters and input layer data to calculate v and joint probability distribution function of h, and use this function value as the initial matrix weight. Probability distribution function as follows,

$$p(h_j = 1|v) = \sigma(c_j + \sum_i w_{ij} v_i) \tag{1}$$

where the sigmoid function is expressed as follows,

$$\sigma = \frac{1}{1 + \exp(-x)} \tag{2}$$

The initial weighted matrix is defined as $w_{ij}(i = 1, 2, \ldots n)$, and the network input data is defined as x, the network output data is defined as $h_{w,\,b}(x)$. In the coding phase, input data x through the mapping function is activated to give u as follows,

$$u = g(w_i x + b_i) \tag{3}$$

where an activation function $g(\cdot)$ is a sigmoid function,

$$g(\cdot) = \frac{1}{1 + e^{-x}} \tag{4}$$

In the decoding stage, the reconstructed signal is,

$$h_{w,b}(x) = g({w_i}^T u + b_{i+1}) \tag{5}$$

3.2.2. Softmax regression

We consider here a Softmax classifier which is the expansion of the logical classifier. The logic classifier is more suitable for some nonlinear classification problems, and it is only suitable for the binary classification problem. The classification results are the categories of probability as output; the final category is determined by the threshold. The Softmax classifier can expand the logic classifier, and have abilities to carry out multi-class classification.

Here we calculate probability value, and compare this result with threshold ϕ, which can be converted into a simple binary classification problem. The expression of the logical function is shown as follows,

$$h_\theta(x) = g\left(\theta^T x\right) \frac{1}{1 + e^{-\theta^T x}} = p(y = 1|x; \theta) \tag{6}$$

where $h_\theta(x)$ is the probability of 1, and θ is the model parameter.

The optimization of the parameters is completed through successive adjustments to minimize the loss function. The loss function is expressed in the following form,

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}}\right] \tag{7}$$

For multiple Softmax classifier, the expansion reads in the form,

$$h_\theta(x^{(i)}) = \begin{bmatrix} p(y^{(i)} = 1|x^{(i)}; \theta) \\ p(y^{(i)} = 2|x^{(i)}; \theta) \\ \vdots \\ p(y^{(i)} = k|x^{(i)}; \theta) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{\theta_j^T x(i)}} \begin{bmatrix} e^{\theta_1^T x(i)} \\ e^{\theta_2^T x(i)} \\ \vdots \\ e^{\theta_k^T x(i)} \end{bmatrix} \tag{8}$$

where the output value is a k dimensional vector of classes, and the model parameters are $\theta_1, \theta_2, \ldots, \theta_k$.

For each category, j outputs calculated probability, which indicates the probability of the data object divided into this category. It has achieved the categorization with the category corresponding to the maximum probability value.

SR is used to sort the features learned by deep sparse autoencoder network, for training set: $\{(x^{(1)}, y^{(1)}), \ldots, (x^{(m)}, y^{(m)})\}$, it has $y^{(i)} \in \{1, 2, \ldots, k\}$. There are $k = 7$ categories in facial emotion recognition problem, are neutral, happiness, anger, sadness, surprise, disgust, and fear. To estimate the probability of belongingness to each category, the function $h_\theta(x)$ (8) is used.

The matrix of parameters of the model is defined as follows,

$$\theta = \begin{bmatrix} \theta_1^T \\ \theta_2^T \\ \vdots \\ \theta_k^T \end{bmatrix} \tag{9}$$

The SR's cost function is given by (10). We add a decay component in the cost function in order to penalize too large values of parameter, and the cost function is rewritten as,

$$J(\theta) = -\frac{1}{m}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} 1\{y^{(i)} = j\} \log \frac{e^{\theta_j^T x^{(i)}}}{\sum_{l=1}^{k} e^{\theta_j^T x^{(i)}}}\right] + \frac{\lambda}{2}\sum_{i=1}^{k}\sum_{j=0}^{n}\theta_{ij}^2 \tag{10}$$

A standard gradient-based optimization method is expressed in the form

$$\theta_j = \theta_j - \alpha \nabla_{\theta_j} J(\theta)(j = 1, 2 \ldots, k) \tag{11}$$

Using the above iterative scheme, the Softmax classification model is optimize.

3.2.3. Overall weight training

Considering the data composed of m samples the minimized performance index reads in the form,

$$J(w, b) = \left[\frac{1}{m}\sum_{i=1}^{m} J(w, b, x^{(i)}, y^{(i)})\right] + \frac{\lambda}{2}\sum_{l=1}^{n_l - 1}\sum_{i=1}^{s_l}\sum_{j=1}^{s_{l+1}} \left(w_{ji}^{(l)}\right)^2 \tag{12}$$

where the first item of the formula is a mean square error term, and the second one describes weighted decay.

Forward conduction equation is employed to calculate of activation value in each layer of network (not including the output layer), so that the following expression holds,

$$a^{(l+1)} = f(w^{(l)} a^{(l)} + b^{(l)}) \tag{13}$$

Recursively calculating residual error, the pertinent formula is given in the form,

$$\delta_i^{(n_l)} = \left(\sum_{j=1}^{s_{l+1}} w_{ji}^l \delta_i^{(l+1)}\right) f'(z_i^{n_l}) \tag{14}$$

For $i = 1, 2 \ldots, m$, The update formulas come in the form,

$$\begin{aligned} \Delta w^{(l)} &= \Delta w^{(l)} + a_j^{(l)} \delta_i^{(l+1)} \\ w^{(l)} &= w^{(l)} - a\left[\left(\tfrac{1}{m}\right)\Delta w^{(l)}\right] + \lambda w^{(l)} \\ \Delta b^{(l)} &= \Delta b^{(l)} + \delta_i^{(l+1)} \\ b^{(l)} &= b^{(l)} - a\left[\tfrac{1}{m}\Delta b^{(l)}\right] \end{aligned} \tag{15}$$

4. Experimental studies

4.1. Data setting

The JAFFE database [44] contains 213 facial expression images of female facial expression corresponding to 10 distinct subjects. Each image is stored at a resolution of 256 * 256 pixels and 8-bit gray level. Each subject in the database is represented with 7 categories of expression (i.e., neutral, happiness, sadness, fear, angry, disgust, and surprise), and each subject has 2–4 images per expression. The sample images of JAFFE are shown in Fig. 3(a).

The CK+ expression database [45] is used as the other experimental samples, which includes more than 100 performers from different regions, with different colors, ages and genders, and contains expression image sequences starting from the neutral emotional state and finishing at the expression apex. The last image from each sequence are selected as sample

(a) JAFFE facial expression library samples (b) CK+ facial expression library samples

Fig. 3. Facial expression library samples.

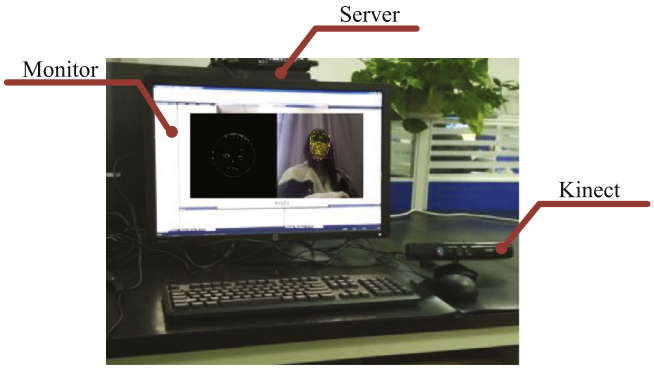

Fig. 4. Emotional computing workstation.

where the expression is at its peak intensity, and the neutral emotional state images are also selected. The sample images of CK+ are shown in Fig. 3(b).

In the experiment that the test data set is the same as the training data set, we use all facial expression images of JAFFE and CK+ as training data set. In the experiment that the test data set is different from training data set, the facial expression images of JAFFE and CK+ were divided into 3 groups, each consisting of a similar number of images.

4.2. Setting

System workflow uses kinect located on the top of the wheeled robot to track facial expression images first, then invokes facial emotion recognition algorithm for feature extraction and emotion recognition, which relies on the affective computing workstations that is as shown in Fig. 4.

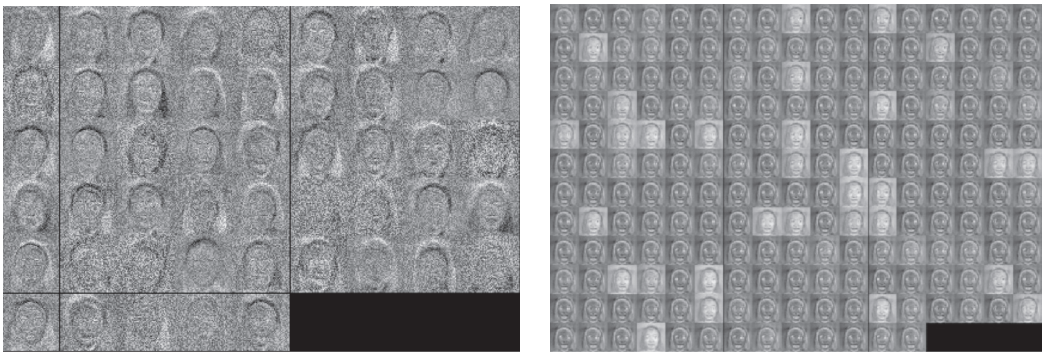

Fig. 5. Visualization of Weights for varying number of nodes in the hidden layer: (a) 50 , and (b) 200 different hidden layer node number.

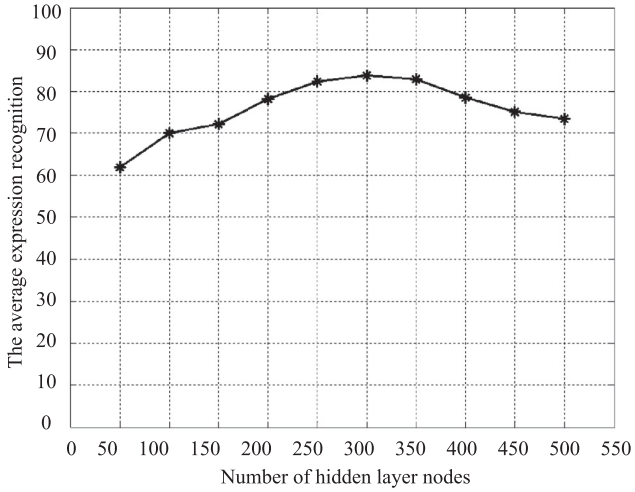

Fig. 6. The influence of the number of hidden layer nodes on the expression recognition rate (for training data set).

4.3. Impact of the number of hidden layer nodes on the performance of recognition

It can be seen from Fig. 5(a) and (b) that the increase of the numbers of hidden layer nodes can increase the recognition rate of emotion, however, too large numbers of hidden layer nodes does not help improve the recognition rate of emotion. Instead it leads to possible over fitting effect. After fine-tune the weights matrix, the recognition rate improves to a certain extent, and offsets the impact of the hidden layer nodes numbers change, as shown in Fig. 6. According to the obtained results, the size of the hidden layer is set to 300. The size of the layer is learned by trial and error. In the series of experiments the number of nodes in the hidden layer changed from 50 to 500, and the highest recognition rate was achieved for 300 neurons in this layer. Refer to Fig. 6.

4.4. Fine-tuning effect on recognition performance

The visualization of underlying characteristics for weights that learned by sparse autoencoder network is designed, and the numbers of neurons node in the hidden layer is set to 140 to obtain an initial image of characteristics' visualization. Visual characteristics weight matrix is shown as follows.

Fig. 7(a) and (b) show the matrices of weights visualization before and after fine-tuning of the weights, it can be seen that the features self-learned by overall network looks more sophisticated after fine-tune the weights to ensure high recognition accuracy. Sparsity has an impact on the recognition rate as illustrated in Fig. 8; the optimal value of sparsity was found to be 0.045 essentially according to Fig. 8. The average expression recognition rate is improved as sparsity parameter gets higher, but it reaches a plateau after the sparsity parameter is set to 0.045, so in the experiment, the sparsity parameter is set to 0.045.

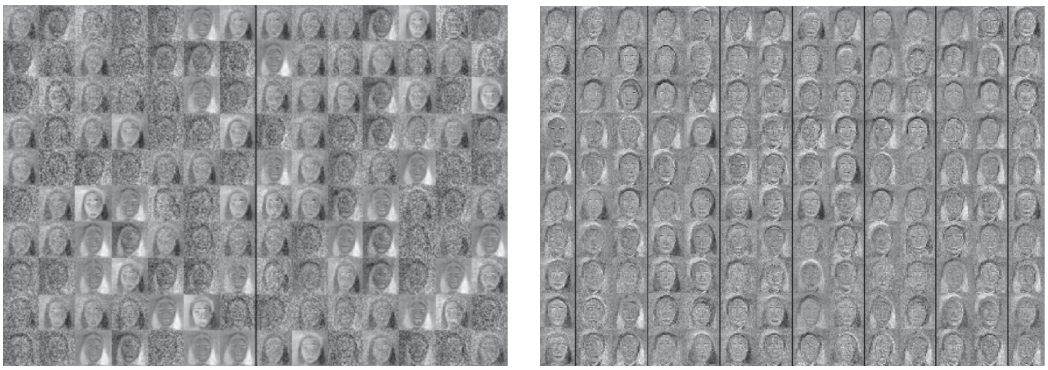

Fig. 7. Weights visualization of the underlying characteristics: (a) before fine tuning, and (b) after fine tuning.

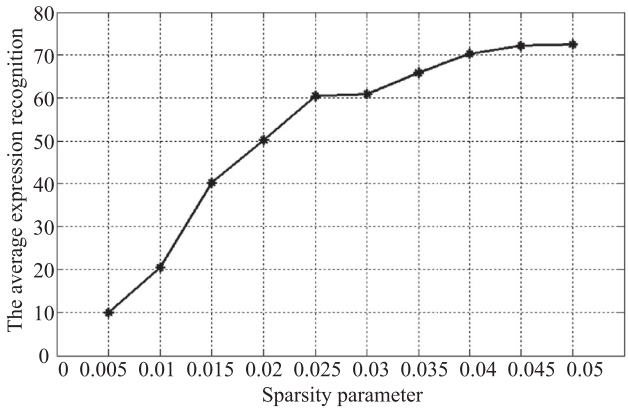

Fig. 8. The influence of sparse parameter and fine-tuning the weights on the rate of facial emotion recognition (for training data set).

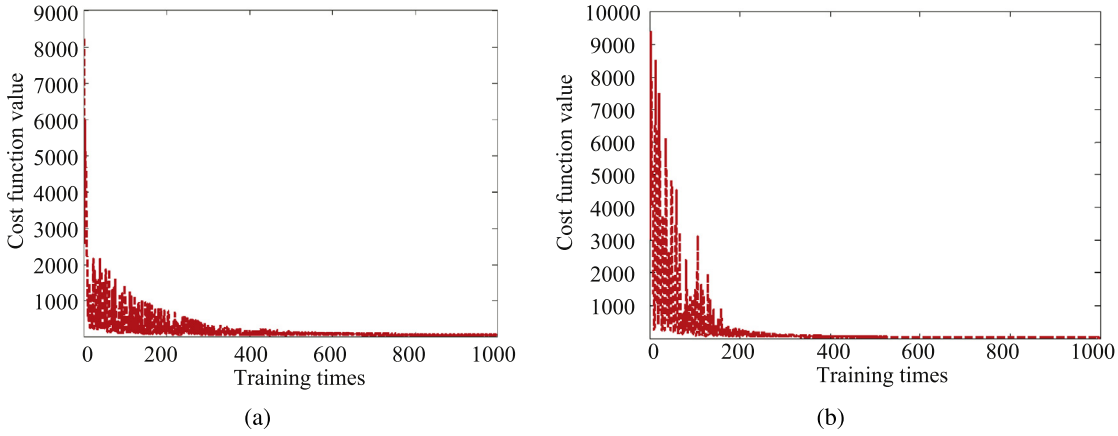

Fig. 9. The convergence of overall cost function: (a) before fine tuning, and (b) after fine tuning.

According to the comparison results shown in Fig. 9(a) and (b), it is obvious that fine-tune can make the overall cost function converge faster.

4.5. Results of recognition

To verify the accuracy of emotion recognition, we conducted two sets of experiments on JAFFE database and CK+ database. In the experiment that the test data set is the same as the training data set, the results are listed in Table. 2,

Table 2
Comparison of emotion recognition experimental results (the training data set is the same as the test data set).

Index	JAFFE		CK+	
	SR	SRDSAN	SR	SRDSAN
Average accuracy(%)	87.79	98.59	100.00	100.00
Training time(s)	0.715242	101.246846	18.402923	163.865084
Test time(s)	0.122398	0.128280	0.125981	0.132169

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.90	0	0	.33	0	.67	0
DI.	.10	.83	0	.07	0	0	0
FE.	.09	0	.75	0	.03	0	.13
HA.	0	0	0	1	0	0	0
NE.	0	0	.03	0	.90	.03	.04
SA.	0	.03	.03	.03	0	.91	0
SU.	.03	0	0	0	.10	0	.87

(a) Confusion matrices of SR on JAFFE

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	1	0	0	0	0	0	0
DI.	0	.93	0	0	0	0	.07
FE.	0	0	.97	0	0	0	.03
HA.	0	0	0	1	0	0	0
NE.	0	0	0	0	1	0	0
SA.	0	0	0	0	0	1	0
SU.	0	0	0	0	0	0	1

(b) Confusion matrices of SRDSAN on JAFFE

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	1	0	0	0	0	0	0
DI.	0	1	0	0	0	0	0
FE.	0	0	1	0	0	0	0
HA.	0	0	0	1	0	0	0
NE.	0	0	0	0	1	0	0
SA.	0	0	0	0	0	1	0
SU.	0	0	0	0	0	0	1

(c) Confusion matrices of SR on CK+

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	1	0	0	0	0	0	0
DI.	0	1	0	0	0	0	0
FE.	0	0	1	0	0	0	0
HA.	0	0	0	1	0	0	0
NE.	0	0	0	0	1	0	0
SA.	0	0	0	0	0	1	0
SU.	0	0	0	0	0	0	1

(d) Confusion matrices of SRDSAN on CK+

Fig. 10. Confusion matrices (the training data set is the same as the test data set).

Table 3
Comparison of emotion recognition experimental results (the training data set and the test data set is different).

Index	JAFFE			CK+		
	SR	CNN [17]	SRDSAN	SR	SCTI [33]	SRDSAN
Group one(%)	70.42	–	84.51	71.07	–	89.31
Group two(%)	85.71	–	94.29	76.54	–	89.51
Group three(%)	68.57	–	88.57	79.63	–	88.27
Average accuracy(%)	74.90	86.74	89.12	75.75	88.70	89.03
Training time(s)	2.702507	–	125.72604	16.396232	–	163.575725
Test time(s)	0.111839	–	0.1025045	0.122803	–	0.133317

Fig. 10 shows the confusion matrices by using SR and SRDSAN on JAFFE database and CK+ database respectively. According to these table and confusion matrices, we can see that the SRDSAN can really learn useful features.

In the experiment that the test data set is different from training data set, the results are listed in Table. 3. As it is shown, the average recognition accuracy of SR is 74.90% in the final test on JAFFE database. Nevertheless, if we use unlabeled training data to train DSAN firstly, and then train the SR model, and we can find that the times of iteration convergence is only 181, and the average accuracy is 89.12%, which is higher than the CNN in [17]. Similarly, the average recognition accuracy of SR is 75.75% in the final test on CK+ database, yet the average accuracy is 89.03% by using SRDSAN which is higher than the traditional method, that is, SCTI [33]. By fusing the SR in deep learning, it can be seen that the features self-learned by overall network looks more sophisticated after fine-tune, and fine-tune makes the overall cost function converge faster, which overcomes the local extrema and gradient diffusion problems. Moreover, this shows that the characteristics learned by our self-learning sparse autoencoder network are more representative than the characteristics in original input data, which is the typical difference between conventional training methods and deep learning training methods. In addition, texture and

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	1	0	0	0	0	0	0
DI.	0	1	0	0	0	0	0
FE.	.10	.10	.70	.10	0	0	0
HA.	0	0	0	1	0	0	0
NE.	.10	0	.10	0	.80	0	0
SA.	0	.20	.10	0	0	.70	0
SU.	0	0	.10	0	.10	0	.80

(a) Confusion matrices of SR on JAFFE

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	1	0	0	0	0	0	0
DI.	0	1	0	0	0	0	0
FE.	.10	0	.60	0	.20	.10	0
HA.	0	0	0	1	0	0	0
NE.	0	0	0	0	1	0	0
SA.	0	0	0	0	0	1	0
SU.	0	0	0	0	0	0	1

(b) Confusion matrices of SRDSAN on JAFFE

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.69	.06	0	0	.25	0	0
DI.	.15	.75	0	.05	0	.05	0
FE.	.06	.13	.75	0	0	.06	0
HA.	0	0	.03	.97	0	0	0
NE.	.17	.03	0	.03	.70	.07	0
SA.	.05	0	0	0	.35	.60	0
SU.	0	0	0	0	.03	0	.97

(c) Confusion matrices of SR on CK+

	AN.	DI.	FE.	HA.	NE.	SA.	SU.
AN.	.69	0	0	0	.31	0	0
DI.	0	.80	0	0	.20	0	0
FE.	0	0	1	0	0	0	0
HA.	0	0	0	1	0	0	0
NE.	0	0	0	0	1	0	0
SA.	0	0	0	0	.40	.60	0
SU.	0	0	0	0	0	0	1

(d) Confusion matrices of SRDSAN on CK+

Fig. 11. Confusion matrices (the training data set and the test data set is different).

Fig. 12. Emotional social robot system.

shape feature changes in the three key parts ROI such as eyebrows, eyes, mouth may reflect changes in facial expression features exactly. Fig. 11 shows the confusion matrices by using SR and SRDSAN on JAFFE database and CK+ database respectively. And according to these confusion matrices, compared with using SR, the recognition accuracy of various expressions has been effectively improved by using SRDSAN.

The proposal SRDSAN is verified in the preliminary application experiments by using our developing ESRS, which includes two mobile robots and a high-performance computer for affective computing, as shown in Fig. 12. It is really an important way to autonomous robots for emotion recognition in HRI. First, Kinect is used for image acquisition, then, the image was conducted on ROI clipping. With that, the image information is applied to the proposed method that is SRDSAN. In the end, facial expression recognition results will be output. The basic experimental results are shown in Fig. 13.

For further research, deep level understanding included emotional intention understanding [4] is being popular in HRI. As the development of the cognitive science, autonomous robots are endowed with higher intelligence, e.g. Pepper [7]. Research topics would be also interesting to extent our multi-robot behavior adaptation mechanisms from emotion and atmosphere [5,6] to intention. According to preliminary application experiments by using the developing ESRS, showing the evidence that the proposed SRDSAN is an effective way to autonomous robots for emotion recognition in HRI.

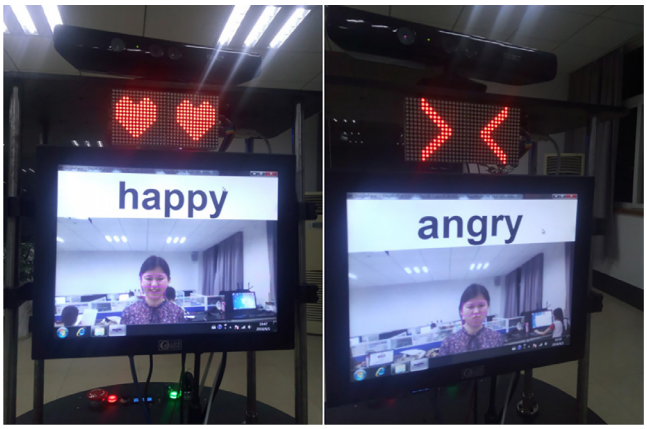

Fig. 13. The preliminary application experiments.

5. Conclusion

In this paper, SRDSAN is proposed for facial emotion recognition to address the problems of learning efficiency and computational complexity, where DSAN is used to extract high-level features and learn facial emotion features, besides, SR is employed to classify facial expression.

According to the experiment results on JAFFE and CK+, by fusing the SR in the proposed method, it can be seen that the features self-learned by overall network looks more sophisticated after fine-tune, and the fine-tune makes the overall cost function converge faster, which overcomes the local extrema and gradient diffusion problems. Consequently, it is clearly that the proposal is an effective method to accomplish facial emotion recognition in HRI and can promote efficiency of the algorithm.

Moreover, basic experiments are carried out in the developing HRI system called ESRS, and the proposal is being extended to emotional social robots for analyzing and understanding human emotions, as well as responding appropriate social behaviors.

Acknowledgements

This work was supported by the National Natural Science Foundation of China under Grants 61603356, 61210011, 61733016, and 61773353, the Hubei Provincial Natural Science Foundation of China under Grant 2015CFA010, the 111 project under Grant B17040, the Fundamental Research Funds for the Central Universities, China University of Geosciences (Wuhan) (No. 201548), and the Wuhan Science and Technology Project under Grant 2017010201010133.

References

[1] A. Aly, A. Tapus, Towards an intelligent system for generating an adapted verbal and nonverbal combined behavior in human crobot interaction, Auton. Robots 40 (2) (2016) 1–17.
[2] M.A.A. Dewana, E. Grangerb, G.L. Marcialisc, R. Sabourinb, F. Rolic, Adaptive appearance model tracking for still-to-video face recognition, Pattern Recognit. 49 (2016) 129–151.
[3] S. An, W. Liu, S. Venkatesh, Face recognition using kernel ridge regression, in: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2007, pp. 1–7.
[4] L.F. Chen, Z.T. Liu, M. Wu, M. Ding, F.Y. Dong, K. Hirota, Emotion-age-gender-nationality based intention understanding in human-robot interaction using two-layer fuzzy support vector regression, Int. J. Soc. Robot 7 (5) (2015) 709–729.
[5] L.F. Chen, Z.T. Liu, F.Y. Dong, Y. Yamazaki, M. Wu, K. Hirota, Adapting multi-robot behavior to communication atmosphere in humans-robots interaction using fuzzy production rule based friend-q learning, J. Adv. Comput. Intell. Intell. Inform. 17 (2) (2013) 291–301.
[6] L.F. Chen, Z.T. Liu, M. Wu, F.Y. Dong, Y. Yamazaki, K. Hirota, Multi-robot behavior adaptation to local and global communication atmosphere in humans-robots interaction, J. Multimodal User Interfaces 8 (3) (2014) 289–303.
[7] S. Calinon, P. Kormushev, D.G. Caldwell, Pepper learns together with children: development of an educational application, in: Proceedings of IEEE-RAS 15th International Conference on Humaniod Robots, Seoul, Korea, 2015, pp. 270–275.
[8] B.J.T. Fernandes, G.D.C. Cavalcanti, T.I. Ren, Face recognition with an improved interval type-2 fuzzy logic sugeno integral and modular neural networks, IEEE Trans. Syst., Man, Cybern.-Part A 41 (5) (2011) 1001–1012.
[9] J. Garcia, R. Salmern, C. Garca, Standardization of variables and collinearity diagnostic in ridge regression, Int. Stat. Rev. 84 (2) (2015) 245–266.
[10] M. Gong, J. Liu, H. Li, A multiobjective sparse feature learning model for deep neural networks, IEEE Trans. Neural Netw. Learn. Syst. 26 (12) (2015) 3263–3277.
[11] C.K. Hsieh, S.H. Lai, Y.C. Chen, An optical flow-based approach to robust face recognition under expression variations, IEEE Trans. Image Process. 19 (1) (2010) 233–240.
[12] G.E. Hinton, R.R. Salakhutdinov, Reducing the dimensionality of data with neural networks, Science 313 (5786) (2006) 504–507.
[13] N. Herndon, D. Caragea, A study of domain adaptation classifiers derived from logistic regression for the task of splice site prediction, IEEE Trans. Nanobioscience 15 (2) (2016) 75–83.
[14] A. Kleinsmith, N. Bianchi-Berthouze, Affective body expression perception and recognition: a survey, IEEE Trans. Affect Comput. 4 (1) (2013) 15–33.

[15] B.K. Kim, J. Roh, S.Y. Dong, Hierarchical committee of deep convolutional neural networks for robust facial expression recognition, J. Multimodal User Interfaces 10 (2) (2016) 1–17.
[16] C. Karyotis, F. Doctor, R. Iqbal, A. James, V. Chang, A fuzzy computational model of emotion for cloud based sentiment analysis, Inf. Sci. (2017). doi:10.1016/j.ins.2017.02.004
[17] A.T. Lopes, E.D. Aguiar, A.F.D. Souza, Facial expression recognition with convolutional neural networks: coping with few data and the training sample order, Pattern Recognit. 61 (2017) 610–628.
[18] B. Leng, Y. Liu, K. Yu, 3D object understanding with 3d convolutional neural networks, Inf. Sci. 366 (2016) 188–201.
[19] T. Maul, T. Maul, T. Maul, Local receptive field constrained deep network, Inf. Sci. 349 (C350) (2016) 229–247.
[20] Q. Mao, Q. Rao, Y. Yu, Hierarchical bayesian theme models for multi-pose facial expression recognition, IEEE Trans. Multimedia 16 (4) (2017) 861–873.
[21] O. Ouyed, M.S. Allili, Feature relevance for kernel logistic regression and application to action classification, in: Proceedings of IEEE International Conference on Pattern Recognition, 2014, pp. 1325–1329.
[22] M.M.A. Rahhal, Y. Bazi, H. Alhichri, Deep learning approach for active classification of electrocardiogram signals, Inf. Sci. 345 (2016) 340–354.
[23] T. Rabie, Training-less color object recognition for autonomous robotics, Inf. Sci. 418–419 (2016) 218–241.
[24] C. Shi, C.-M. Pun, 3d multi-resolution wavelet convolutional neural networks for hyperspectral image classification, Inf. Sci. 420 (2017) 49–65.
[25] D.Y.Y. Sim, C.K. Loo, Extensive assessment and evaluation methodologies on assistive social robots for modelling human-robot interaction-a review, Inf. Sci. 301 (2015) 305–344.
[26] L. Shao, Z. Cai, L. Liu, Performance evaluation of deep feature learning for RGB-d image/video classification, Inf. Sci. 385 (2017) 266–283.
[27] R.V. Sharan, T.J. Moir, Robust acoustic event classification using deep neural networks, Inf Sci (Ny) 396 (2017) 24–32.
[28] Y. Sun, Y. Chen, X. Wang, Deep learning face representation by joint identification-verification, Adv. Neural Inf. Process. Syst. 27 (2014) 1988–1996.
[29] M. Tkal, A. Odi, A. Koir, The impact of weak ground truth and facial expressiveness on affect detection accuracy from time-continuous videos of facial expressions, Inf. Sci. 249 (16) (2013) 13–23.
[30] M.Z. Uddin, M.M. Hassan, A. Almogren, Facial expression recognition utilizing local direction-based robust features and deep belief network, IEEE Access 5 (2017) 4525–4536.
[31] P.Y. Wu, C.C. Fang, J.M. Chang, S.Y. Kung, Cost-effective kernel ridge regression implementation for keystroke-based active authentication system, IEEE Trans. Cybern. (2017).
[32] Y. Wang, X. Wang, W. Liu, Unsupervised local deep feature for image recognition, Inf. Sci. 351 (2016) 67–75.
[33] J. Yi, X. Mao, L. Chen, Facial expression recognition considering individual differences in facial structure and texture, IET Comput. Vision 8 (5) (2014) 429–440.
[34] S.Y. Xie, H.F. Hu, Facial expression recognition with FRR-CNN, Electron. Lett. 53 (4) (2017) 235–237.
[35] H. Xue, Y. Zhu, S. Chen, Local ridge regression for face recognition, Neurocomputing 72 (4) (2009) 1342–1346.
[36] Y. Xu, Z. Li, B. Zhang, Sample diversity, representation effectiveness and robust dictionary learning for face recognition, Inf. Sci. 375 (2017) 171–182.
[37] Y. Xu, Q. Zhu, Z. Fan, Using the idea of the sparse representation to perform coarse-to-fine face recognition, Inf. Sci. 238 (7) (2013) 138–148.
[38] X. Yang, W. Liu, D. Tao, Canonical correlation analysis networks for two-view image recognition, Inf. Sci. 385–386 (2017) 338–352.
[39] M. Yang, P. Zhu, F. Liu, Joint representation and pattern learning for robust face recognition, Neurocomputing 168 (C) (2015) 70–80.
[40] L. Zhang, D. Tjondronegoro, Facial expression recognition using facial movement features, IEEE Trans. Affect. Comput. 2 (4) (2011) 219–229.
[41] L. Zhang, P.N. Suganthan, A survey of randomized algorithms for training neural networks, Inf. Sci. 364 (2016) 146–155.
[42] T. Zhang, W. Zheng, Z. Cui, A deep neural network-driven feature learning method for multi-view facial expression recognition, IEEE Trans. Multimedia 18 (12) (2016) 2528–2536.
[43] X. Zhao, X. Shi, S. Zhang, Facial expression recognition via deep learning, IETE Tech. Rev. 32 (5) (2015) 347–355.
[44] The japanese female facial expression (JAFFE) database, 1998, http://www.kasrl.org/jaffe.html.
[45] Cohn-Kanade, (CK and CK+) database download site, 2000, http://www.consortium.ri.cmu.edu/data/ck/.