# Facial Expression Recognition Based on Arousal-Valence Emotion Model and Deep Learning Method

Yong Yang
Chongqing Key Laboratory of Computational and Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
yangyong@cqupt.edu.cn

Yue Sun
Chongqing Key Laboratory of Computational and Intelligence
Chongqing University of Posts and Telecommunications
Chongqing, China
suny0702@163.com

***Abstract*—The traditional facial emotion recognition method is classifying basic emotions. But, basic emotions theory is limited to express subtle and disparate emotion. So this paper uses the arousal-valence continuous emotion space model, which can enrich emotion expression. The arousal reflects emotional intensity, and the valence indicates positive and negative emotion. The arousal and valence all have the value in the same range, which is between -1 and 1. In the experiments, it uses convolutional neural network (CNN) in the pre-trained models and support vector regression(SVR). In this model, CNN works as a trained feature extractor and SVR is adopted to train and predict the values of the arousal and valence. Through the predicted values it can be predicted the facial emotion. The contrast experimental results show that the proposed method can get better recognition result than the traditional methods.**

***Keywords — facial emotion recognition, arousal-valence emotion dimensions, convolution neural network(CNN), support vector regression(SVR).***

## I. INTRODUCTION

With the development of artificial intelligence, facial expression recognition has been extensively studied in the human-computer interaction. Facial expression is the human's main way to express emotion. Psychologist Mehrabian defined the well-known 7%-38%-55% criterion for emotion expression, that is, emotion is expressed by 7% language, 38% voice and 55% facial expression [1]. In recent years, facial expression recognition is applied in many areas, such as human-computer interaction, driver fatigue monitor and patient care in hospital.

Basic emotion theory was proposed by Ekman included six categories, such as anger, happy, sad, surprise, disgust and fear [2]. Basically, this theory has largely focused on classification. These basic emotion theory has limited ability to express emotions. On the one hand, the basic theory has its limitations. On the other hand, classification can't distinguish the relationship between different emotions. Most importantly, it can't express the intensity of expression. Recently, many researchers in the field of affective computing resort to subtle and continuous emotion model [3]. Russell proposed the continuous common emotional models include valence and arousal emotion model [4], Davis proposed pleasure, arousal, power emotional model [5], Mehrabian proposed the pleasure, arousal, dominance emotional dimension model [6] and so on. Arousal-valence model is one of these models, it is widely used and can effectively describe a person's emotional change.

Within the arousal-valence(A-V) model of human emotion, where arousal reflects emotional intensity and valence indicates positive vs. negative emotions [7]. The value of arousal and valence has the same range, the range is between -1 and 1.

In addition, the paper proposed to use convolutional neutral network (CNN) [8] to extract face features. Traditional face feature extraction is a tedious and time-consuming process, it is easy to lose face feature characterizing details. While convolutional neutral network can automatically evaluate the optimized features. Then the CNN features combine Support Vector Regression(SVR) to predict facial expression. CNN-SVR recognition method is proposed for facial expression recognition. This proposed method uses features based on the CNN model, and predicts facial expression using the SVR regressor.

## II. CNN-SVR EMOTION RECOGNITION METHOD

The proposed method is CNN-SVR, which is designed by the CNN features combine Support Vector Regression(SVR) to predict facial expression. Normally, a Convolutional Neural Network's last output layer is for classifying. It's inapposite in the arousal-valence model. In this paper, the CNN is taken as the invisible feature extractor, the SVR is used to predict the emotion's dimension, which are respectively arousal and valence. The CNN model is based on VGG structure, Fig.1 shows the proposed CNN-SVR recognition method.

### *A. The Proposed CNN-SVR Method*

In this paper, CNN is regarded as feature extractor, it is like biological neural network, and it can avoid the traditional complicated process of feature extraction and data reconstruction. In addition, in the arousal-valence model, the value of arousal and valence is continuous, in order to predict the value of arousal and valence, SVR regressor is used in this experimemt.

The proposed CNN-SVR method is shown in Fig.1.

978-1-5386-6751-4/18/$31.00 ©2018 IEEE

Dalian, China
September 28-30, 2018

Firstly, the normalized images should deliver to the input layer of the pre-trained CNN, CNN learned the image features after a series of convolution process, pooling prcess and full connection process, then the outputs of layer in the model were fed as features to the SVR.

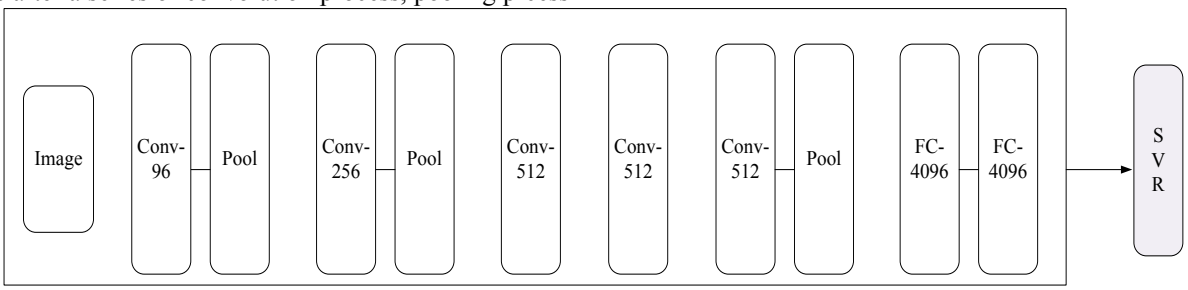


Fig.1. The proposed CNN-SVR method

On the other hand, arousal and valence model is used in this paper, in order to predict the value of the arousal and valence, regressor(SVR) should be used in this experiment. Support vector regression is the regression method, SVR likes other machine learning, has simple structure, through the training, it solves the function fitting problems. Besides, SVR has high performance and strong ability of generalization, avoiding dimension curse.

Since CNN has already achieved better performance in face recognition[9], the proposed method CNN-SVR focus on CNN features extraction and regression, which can yield better results.

### *B. Convolutional Netural Network*

Lecun Y and others [10] proposed the convolutional netural network(CNN), it is the first successful learning algorithm of training multi-layer network. CNN has five basic structures, they are input layer, convolution filtering layer, pooling layer, full connection layer and output layer. As shown in Fig.2, it is a typical CNN model.

CNN is one of the deep learning models, it can be view as an automatic extractor. The feature extractor contains feature map layers and retrieves discriminating features from the raw images via three main layers: convolution filtering layers, pooling layers, full connection layers. Specific operation as follows:

- Convolution filtering layers: The convolutional layer is the core building block of a CNN. It forms the basis of the CNN and performs the core operations of training and consequently firing the neurons of the network. The layer's parameters consist of a set of learnable filters (or kernels), which have a small receptive field. Convolutional layer is actually feature extraction layer, convolution operation can enhance original signal characteristics, at the same time, it also can reduce noise.
- Pooling layers: Its function is to progressively reduce the spatial size of the representation to reduce the amount of parameters and computation in the network, and hence to also control overfitting. Pooling layer using the max operation to calculate every depth slice of the input and resizes it spatially, Maxpooling is done by applying a max filter to non-overlapping sub-regions of the initial representation. Pooling layer is used for further feature extraction, it can be viewed as a fuzzy filter.
- Full connection layers: It can be seemed as the multilayer perceptron's hidden layers. Every neuron from the previous layer is connected to every neuron on the next layer, and each connection has its own weight. This is a totally general purpose connection pattern and makes no assumptions about the features in the data.

### *C. Pre-trained Model*

The paper used VGG-M models[11] to extract face features. It is a pre-trained models from MatConvNet which is MATLAB toolbox implementing Convolutional Neural Network(CNN) for computer vision applications. The model contains five convolutional layers and three full connected layers. Table I has represented the model details.

TABLE I. VGG-MODEL

| Model | Conv1 | Conv2 | Conv3 | Conv4 |
|---|---|---|---|---|
| VGG-M | 97×7×7 | 256×5×5 | 512×3×3 | 512×3×3 |
| | st.2,pad 1 | st.2,pad 1 | st.1,pad 1 | st.1,pad 1 |
| | LRN,x2 pool | LRN,x2 pool | - | - |
| VGG-M | **Conv5** | **Full6** | **Full7** | **Full8** |
| | 512×3×3 | 4096 dropout | 4096 dropout | 1000 softmax |
| | st.1,pad 1 | | | |
| | x2 pool | | | |

Among the convolutional layers' 3 sub-rows, the first sub-row means the number of convolution filters and corresponding field size as " num×size×size"; the following sub-row specifies the convolution stride("st.") and spatial padding("pad"); the last sub-row indicates if Local Response Normalisation (LRN)[12] is applied, and the max-pooling downsampling factor. In 3 full connection layers, the first and the second are regularised with dropout[12], the third layer can be seemed as a classifier. But in this paper, we didn't use the last full connection layer, it is replaced by SVR for regression.

### *D. Support Vector Regression*

A version of support vector machine SVM for regression was proposed in 1996 by Vladimir N. Vapnik and others[13]. It is called support vector regression(SVR). SVM is mainly aimed to solve classifying data, while SVR is produced by support vector classification relies on a subset of the training data, because the cost function for building the model ignores any training data close to the model prediction. It can be applied to regression problems.

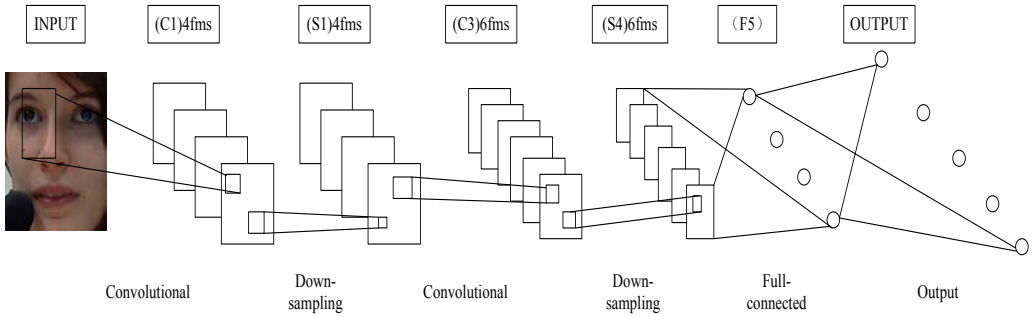


Fig.2. A typical CNN model.

## III. EXPERIMENTS

Experiments is conducted on the AVE2013[14], RECOLA [15] and NVIE[16] datasets in this paper.

Face detection used Haar feature classifier with OpenCV. The image preprocessing method is provided by [17]. CNN-M model requires images to be transformed to a fixed size (224×224). All raw images are sent into CNN and learned the optimized features. In the experiment the model of CNN we used is from MATCONVNET [10], MATCONVNET is a MATLAB toolbox.

In order to reduce the dimensions of the feature space, principal components analysis(PCA) is applied to feature dimension reduction. Meanwhile, PCA principal component contributor rate is set to 0.99. The experiment dataset is divided into training set and test set, training set and test set respectively take up 90% and 10%. The proposed experiment process is shown in Fig.3.

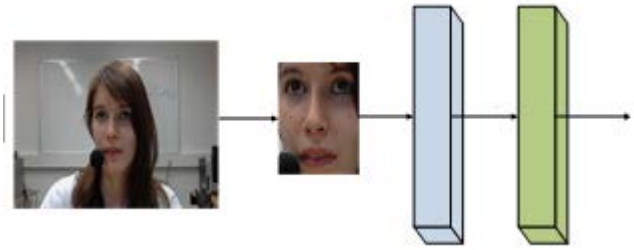

Fig.3. The proposed experiment process

### *A. Experimental Evaluation*

There are three methods to evaluate the experiment in this paper. Respectively, they are Root Mean Square Error(RMSE), Pearson Correlation Coefficient(COR) and Mean Absolute Error(MAE). Assumed there are two n-dimensional variables X, Y, the definition of RMSE, COR and MAE as follow.

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - Y_i)^2} \tag{1}$$

$$COR = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{(\sum X^2 - \frac{(\sum X)^2}{N})(\sum Y^2 - \frac{(\sum Y)^2}{N})}} \tag{2}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|X_i - Y_i\right| \tag{3}$$

The COR denotes the correlation between predictions and the eventual outcomes. The COR ranges from −1 to 1. A value of 1 implies that a linear equation describes the relationship between predictions and the eventual outcomes perfectly. The RMSE makes an excellent general purpose error metric for numerical predictions. Compared to the similar Mean Absolute Error, RMSE amplifies and severely punishes large errors. Its values closer to zero are better. The MAE is a quantity used to measure how closer predictions are to the eventual outcomes. And the MAE is adopted the same evaluation metrics, the values closer to zero are better.

### *B. Compared Experiments And Experimental Results*

To evaluate the effect of the method of CNN-SVR, the traditional feature extraction Gabor, LBP and LPQ are compared. The experimental results are shown in Table II, Table III and Table IV.

TABLE II. EXPERIMENTAL RESULTS OF AROUSAL AND VALENCE ON AVE2013

| AVE Database | A | | |
|---|---|---|---|
| | **RMSE** | **COR** | **MAE** |
| **CNN+SVR** | 0.120 | 0.521 | 0.096 |
| **Gabor+SVR** | 0.158 | 0.273 | 0.124 |
| **LBP+SVR** | 0.163 | 0.259 | 0.130 |
| **LPQ+SVR** | 0.156 | 0.310 | 0.12 |
| **AVE Database** | **V** | | |
| **CNN+SVR** | 0.079 | 0.472 | 0.052 |
| **Gabor+SVR** | 0.100 | 0.399 | 0.077 |
| **LBP+SVR** | 0.097 | 0.425 | 0.074 |
| **LPQ+SVR** | 0.107 | 0.360 | 0.083 |

TABLE III. EXPERIMENTAL RESULTS OF AROUSAL AND VALENCE ON RECOLA

| Recola Database | A | | |
|---|---|---|---|
| | RMSE | COR | MAE |
| CNN+SVR | 0.137 | 0.582 | 0.109 |
| Gabor+SVR | 0.161 | 0.331 | 0.148 |
| LBP+SVR | 0.163 | 0.259 | 0.130 |
| LPQ+SVR | 0.194 | 0.226 | 0.163 |
| Recola Database | V | | |
| CNN+SVR | 0.097 | 0.577 | 0.081 |
| Gabor+SVR | 0.127 | 0.301 | 0.101 |
| LBP+SVR | 0.097 | 0.425 | 0.074 |
| LPQ+SVR | 0.147 | 0.190 | 0.116 |

TABLE IV. EXPERIMENTAL RESULTS OF AROUSAL AND VALENCE ON NVIE

| NVIE Database | A | | |
|---|---|---|---|
| | RMSE | COR | MAE |
| CNN+SVR | 0.330 | 0.601 | 0.279 |
| Gabor+SVR | 0.357 | 0.544 | 0.289 |
| LBP+SVR | 0.368 | 0.519 | 0.292 |
| LPQ+SVR | 0.367 | 0.517 | 0.287 |
| NVIE Database | V | | |
| CNN+SVR | 0.654 | 0.749 | 0.398 |
| Gabor+SVR | 0.558 | 0.645 | 0.461 |
| LBP+SVR | 0.511 | 0.721 | 0.404 |
| LPQ+SVR | 0.522 | 0.697 | 0.411 |

### *C. Experimental Analysis*

As experimental results show in TABLE II-IV, the proposed method can get lower RMSE and MAE while higher COR. For example, in the TABLE II, RMSE is reduced 4.23%-4.65%, and MAE is reduced 2.45%-5.4%, meanwhile COR is increased 7.99%-10.18%. The COR has the obvious improvement. The results indicate that the proposed method is better than traditional method (Gabor, LBP, LPQ). From the contrast experiment, CNN features are better than the other three traditional methods, whenever in RMSE, COR or MAE, CNN features have obvious advantages. The results indicate that the proposed CNN-SVR method is promising regression method due to two properties: Primarily, CNN has the independent ability of expression in corresponding layer. The CNN features avoid the computationally expensive process of traditional method and can implicitly obtain more abstract image feature representation, avoiding explicit process of feature extracting in traditional method. It can input the pixel of image directly, learning data independently through training images. Secondly, SVR has good performance on fitting and prediction, it also can optimize the objectives. SVR adopts the kernel function to avoid the risk of overfitting, it trains the CNN features and gets the reliable forecast. Above all, the experiment gets the desirable results.

## IV. CONCLUSION

Facial expression recognition based on arousal-valence emotion model is currently rare in domestic research. The deep learning is adopted in this paper. It has a significant advantage in facial expression recognition. The future research will focus on softmax layer, the last layer of CNN will be replaced with a projection layer (to perform regression).

## REFERENCES


[1] Mehrabian, A. (2008). Communication without words. Communication Theory, 193-200.

[2] Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. Journal of personality and social psychology, 17(2), 124.

[3] Gunes H, Pantic M. Automatic, Dimensional and Continuous Emotion Recognition[J]. International Journal of Synthetic Emotions, 2010, 1(1):68-99.

[4] Russel, J. A. (1980). Acircumplexmodelofafect. Journal of Personalityand.

[5] Davitz, J. R. (1969). The language of emotion. Academic Press.

[6] Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. Journal of research in Personality, 11(3), 273-294.

[7] Russell , J. A., "A complex model of affect," J. Personality Social Psychology, vol. 39, pp. 1161–1178, 1980.

[8] LeCun, Y., et al., Backpropagation applied to handwritten zip code recognition. Neural computation, 1989. 1(4): p.541-551.

[9] He R, Wu X, Sun Z, et al. Wasserstein CNN: Learning Invariant Features for NIR-VIS Face Recognition[J]. 2017.

[10] LECUN Y, BOSER B, DENKER J S, et al. Backpropagation applied to handwritten zip code recognition [J]. Neural computation, 1989,1 (4): 541-551.

[11] Chatfield K, Simonyan K, Vedaldi A, et al. Return of the Devil in the Details: Delving Deep into Convolutional Nets[J]. Computer Science, 2014.

[12] Krizhevsky A, Sutskever I, Hinton G E. ImageNet classification with deep convolutional neural networks[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2012:1097-1105.

[13] Drucker, Harris; Burges, Christopher J. C.; Kaufman, Linda; Smola, Alexander J.; and Vapnik, Vladimir N. (1997); "Support Vector Regression Machines", in Advances in Neural Information Processing Systems 9, NIPS 1996, 155–161, MIT Press.

[14] Valstar, M., Schuller, B., Smith, K., Eyben, F., Jiang, B., & Bilakhia, S., et al. (2013). AVEC 2013: the continuous audio/visual emotion and depression recognition challenge. ACM International Workshop on Audio/visual Emotion Challenge (pp.3-10).

[15] Ringeval, F., Sonderegger, A., Sauer, J., & Lalanne, D. (2013). Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (pp.1-8).

[16] Wang S, Liu Z, Lv S, et al. A Natural Visible and Infrared Facial Expression Database for Expression Recognition and Emotion Inference[J]. IEEE Transactions on Multimedia, 2010, 12(7):682-691.

[17] Shih, F. Y., & Chuang, C. F. (2004). Automatic extraction of head and face boundaries and facial features. Information Sciences, 158, 117-130.