ICIC Express Letters
Part B: Applications
Volume **7**, Number **9**, September **2016**

 ISSN 2185-2766
pp. **2055–2060**

FACIAL EXPRESSION RECOGNITION USING DEEP NEURAL NETWORK AND DECISION FUSION

Xiuyou Wang and Dongqing Xu

School of Computer and Information Engineering
Fuyang Teachers College
No. 100, Qinghe Rd., Fuyang 236037, P. R. China
xywangfync@sina.com

Received March 2016; accepted June 2016

Abstract. *In this paper we study the facial expression recognition using both low level visual features and high level semantic rules. First, the facial landmark points are localized by Active Shape Model. The key expression regions are then extracted. Second, the Local Binary Pattern features are extracted and a deep neural network is trained by Restricted Boltzmann Machine. Third, the output of the neural network is used for semantic inference, and fuzzy inference system is adopted to implement the high level decision system. Finally, experiments are carried out on the Japanese Female Facial Expression Database. Results show that the proposed system constantly provides higher recognition rates over two traditional methods.*
Keywords: Expression recognition, Deep learning, Semantic inference

1. **Introduction.** Affective Computing (AC) is an important research field in Artificial Intelligence (AI). The motivation of AC is to enable computers to understand human emotions and behaviours at a high level [1, 2, 3]. Facial Expression Recognition (FER), as a sub-field of AC, has drawn many researchers from various backgrounds. It is an interdisciplinary field that involves computer science, electronic engineering, psychology and philosophy.

Traditional expression recognition studies focused on several discrete emotion classes, such as happiness, sadness, and anger. These basic emotion types played an important role in the initial studies [4, 5, 6]. The universal behaviour cues can be revealed in these basic emotions.

With the development of machine learning techniques and psychology theories, the continuous emotion recognition problem has been paid more and more attentions to. Arousal-Valence (AV) dimensional model is one of the most popular emotion models. Unlike the discrete emotion types, the AV space allows us to analyse any type of emotions in a continuous space.

There are many attempts of recognizing emotional dimensions in facial expressions [7, 8]. Most of them follow a bottom-up approach that only considers the texture and local features of facial images. The semantic information of facial expressions is overlooked. Few studies have considered the semantic meanings of facial Action Units (AU). The relative movement of facial components, such as eyebrow movement, nose movement, and mouth movement, are not modelled at a semantic level.

In this paper, we propose to use a two-stage approach to model the facial expressions. We use DBN to model the local facial features and we use semantic inference rules to model the high level facial action units. The system flowchart is shown in Figure 1. The advantages of our work are two-fold: i) the deep network is very effective in modelling the local variance and provides a strong generalization ability; ii) the inference rules are

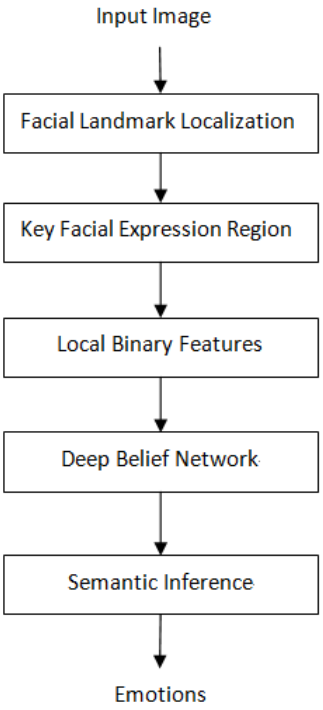

FIGURE 1. System flowchart of the proposed two-stage approach

independent to the local image features and enable our system to support any type of emotions.

The rest of the paper is organized as follows: Section 2 gives the landmark detection method; Section 3 provides the Local Binary Patten (LBP) feature analysis and Deep Belief Network (DBN) model; Section 4 gives the inference rules for high level decision in the AV dimension space; experimental results are given in Section 5; finally, the conclusions are drawn in Section 6.

2. **Key Facial Expression Region.** Facial landmark localization [9] is the first step to process facial images in our system. Active Shape Model (ASM) [10, 11, 12] is adopted to detect the key points in a face, such as eye corners, tip of nose and mouth corners.

Active shape models are statistical models that iteratively deform themselves to fit the shape of objects. It was proposed by T. Cootes et al. [10], and the shape model was defined as a point distribution model. ASM consists of the following two steps: i) generate a shape by searching for a better position in the neighbouring area; ii) conform the new shape to the point distribution model which is a pre-learned probabilistic model.

Each facial landmark is a node on the shape, and the ith shape in the ASM model can be represented as [11]:

$$\mathbf{x}_i = (x_{i,0}, y_{i,0}, x_{i,1}, y_{i,1}, \ldots, x_{i,n-1}, y_{i,n-1})^T \tag{1}$$

Define the similarity of two shapes of facial landmarks as the energy function:

$$E_j = (\mathbf{x} - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j)^T \mathbf{W} (\mathbf{x}_i - M(s_j, \theta_j)[\mathbf{x}_j] - \mathbf{t}_j) \tag{2}$$

where M is the rotation operation defined by angle θ and scaling s. $\mathbf{W}$ is the weight. $\mathbf{t}_j$ is the translation mapping $\mathbf{x}$ onto $M(s_j, \theta_j)[\mathbf{x}_j]$. By optimizing the energy function, we can locate the facial landmarks.

After we localize the landmarks, we can extract key expression regions by three bounding boxes: eyebrow region, nose region and mouth region, as shown in Figure 2. These three regions are closely related to facial expressions. The semantic meanings of the movement of the facial components are based on the three bounding boxes.

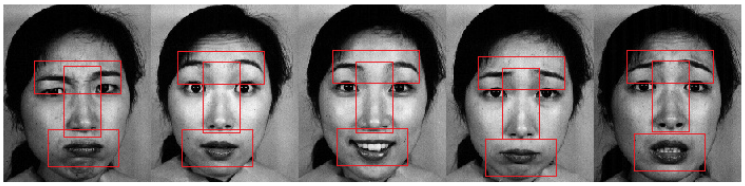

FIGURE 2. A depiction of key facial expression regions

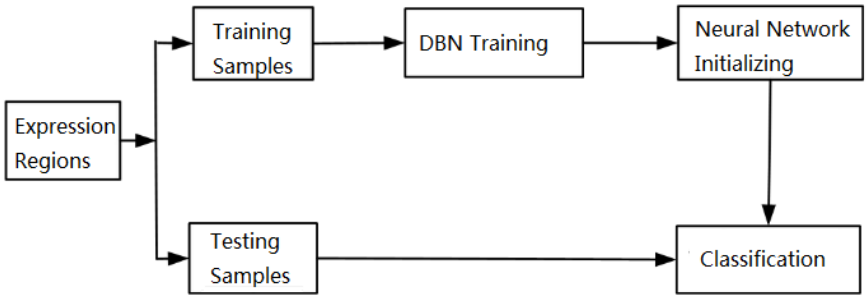

FIGURE 3. Flowchart of the neural network classification of arousal-valence levels of key expression regions

3. **Feature Extraction and Modelling.** The extracted facial expression regions are further analysed by LBP features and DBN for the arousal and valence levels. The facial image is labelled by different valence levels and arousal levels according to the expression categories. The arousal dimension is divided into four scales a_1, a_2, a_3, a_4, and the valence dimension is also divided into four scales, a_1, a_2, a_3, a_4. Therefore, the low level feature process can be solved by classification in AV dimensional space.

Local binary pattern is an effective feature for vision related classifications. LBP is the particular case of the Texture Spectrum Model. It is created by the following steps [13].

Step 1) Divide the expression regions into smaller rectangles.

Step 2) Compare the neighbouring eight pixels around each pixel in a rectangle.

Step 3) When the center pixel value is larger, set the feature value to zero; otherwise, one. This provides the simple binary property.

Step 4) Compute histogram of the binary values.

Step 5) Concatenate and normalise the histogram values.

LBP feature is then used to form high level representations through deep neural networks. The Restricted Boltzmann Machine (RBM) is used to initialize the neural network, and the multi-layer neural network is used to classify the arousal-valence labels of each expression region, as shown in Figure 3.

Deep Belief Network consists of multiple RBMs [14]. The structure of RBM is shown in Figure 4. RBM is a typical neural network. Its visible layer s and hidden layer h are connected and the nodes in the same layer are not connected. Therefore, all the hidden nodes are conditionally independent to each other.

$$p(h|v) = p(h_1|v) \dots p(h_n|v) \tag{3}$$

The training process of DBN involves two steps [14], the pre-training and the fine-tuning. In the pre-training, the energy function between the visible layer and the hidden layer is represented as:

$$E(s,h) = -\sum_{i=1}^{S}\sum_{j=1}^{H} w_{ij}s_ih_i - \sum_{i=1}^{S} b_is_i - \sum_{j=1}^{H} a_jh_j \tag{4}$$

where w is the weight and a, b are the offsets.

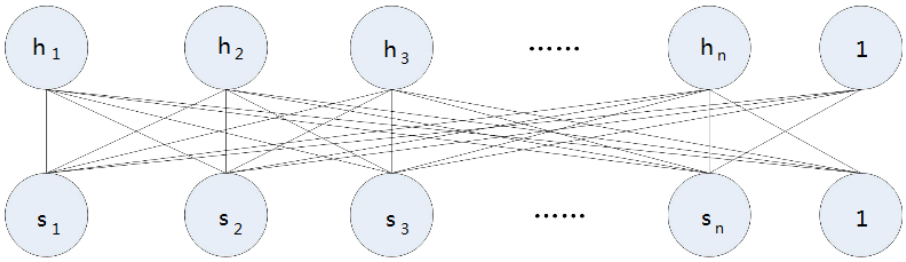

FIGURE 4. A depiction of the RBM structure

In the fine-tuning process the Back Propagation (BP) algorithm is adopted for minimizing the classification error.

4. **Semantic Inference.** Six expression types are considered in the semantic inference, namely anger, joy, sadness, surprise, disgust and fear. However, the supported expression types are not limited to these basic emotions. The semantic vector $\boldsymbol{f}$ is constructed from the three expression regions, as shown in Equation (5). The DBN outputs the arousal-valence levels a and v of the expression regions.

$$\boldsymbol{f} = [a_1, s_1, a_2, s_2, a_3, s_3] \tag{5}$$

where a_1 and s_1 are related to eyebrow region, a_2 and s_2 are related to nose region, a_3 and s_3 are related to mouth region. The expression type can be inferred from the semantic vector $\boldsymbol{f}$, using the rules shown in Table 1.

TABLE 1. Arousal and valence scales of expression regions under various emotions

Emotions	a_1	s_1	a_2	s_2	a_3	s_3
Anger	2	−2	2	−2	2	−2
Joy	1	1	1	1	1	1
Sadness	−1	−1	−1	−1	−1	−1
Surprise	2	0	2	0	2	0
Disgust	0	−2	0	−2	0	−2
Fear	1	−1	1	−1	1	−1
Neutrality	0	0	0	0	0	0

Since the facial expression is influenced by individual personality, culture and context situations, the semantic inference rules need to be further adapted. When the AV dimensions of one expression region are inconsistent with the others, we use fuzzy inference to correct the error.

Let A be the fuzzy set, and the membership function is defined as:

$$\mu_A(x) = 1 - \frac{|x - x^*|}{\delta} \epsilon(\delta - |x - x^*|) \tag{6}$$

where $\epsilon(x)$ is the step function, x is the arousal or valence scale, x^* is the mean value under a specific emotion type, parameter δ is set empirically. The fuzzy inference rules can be changed according to different target emotions. In this paper we adopt the Takagi-Sugeno (TS) implementation which is a non-linear model and time variant model. Based on TS model we implement an accurate inference model consisting of a set of IF-THEN fuzzy rules. The fuzzy rules are converted from the facial actions under different expressions.

5. **Experimental Result.** The Japanese Female Facial Expression Database (JAFFE) [15] is used to verify our expression recognition system. Ten-fold cross validation is adopted. The dataset is divided into ten parts randomly, and the training to testing ration is set to 9. Ten percent of the facial images are used for testing and ninety percent of the facial images are used for training. The averaged recognition rate is used to evaluate the performance of the facial expression system. Examples of the expression data are shown in Figure 5.

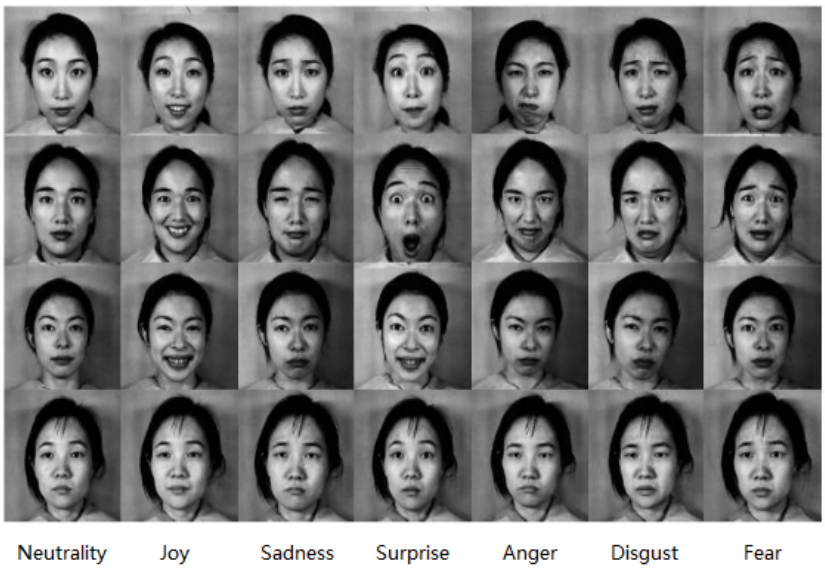

FIGURE 5. Facial expression images used in the experiment

In order to compare the results, we adopt two traditional methods. The first method uses LBP feature and Support Vector Machine (SVM) which is widely applied in machine learning applications. The second method uses LBP features and DBN classifier, which involves deep neural network.

As shown in Table 2, we can see that the proposed system constantly provides higher recognition rates over the two traditional methods, "LBP + SVM" and "LBP + DBN". The proposed algorithm makes use of the semantic rules, and is more robust to random influences in local visual features. The SVM method may be superior under small sample size, but the extendibility is not guaranteed. The DBN method alone can bring good performance; however, it is still affected by the rotations and the landmark detection errors.

TABLE 2. Comparison of expression recognition results using three methods

Algorithms	anger	joy	sadness	surprise	disgust	fear	neutrality
LBP + SVM	93.2%	88.4%	78.6%	79.8%	80.3%	88.2%	75.7%
LBP + DBN	96.1%	90.2%	81.9%	85.8%	82.1%	91.0%	77.5%
Proposed algorithm	97.1%	92.3%	88.3%	89.1%	85.4%	92.3%	81.5%

6. **Conclusions.** In this paper semantic-based inference algorithm is used to adapt the output of low-level visual features for automatic facial expression recognition. The variances in local visual features are handled by DBN, and the learning procedure relies on the successful initialization by Restricted Boltzmann Machine. The output of each key expression region indicates the arousal and valence levels, and the fuzzy rules applied on these AV levels are based on psychology theories which can be extended to any type of

emotions. Recognition results have shown a promising performance compared with another two traditional methods. In the future work, our emotion model may be extended to the control dimension and we will further improve the recognition performance.

REFERENCES

[1] F. Ren and C. Quan, Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: An application of affective computing, *Information Technology and Management*, vol.13, no.4, pp.321-332, 2012.

[2] P. Zimmermann, S. Guttormsen and B. Danuser, Affective computing: A rationale for measuring mood with mouse and keyboard, *International Journal of Occupational Safety and Ergonomics*, vol.9, no.4, pp.539-551, 2003.

[3] C. Huang, R. Liang, Q. Wang, J. Xi, C. Zha and L. Zhao, Practical speech emotion recognition based on online learning: From acted data to elicited data, *Mathematical Problems in Engineering*, 2013.

[4] C. Zou, C. Huang, D. Han and L. Zhao, Detecting practical speech emotion in a cognitive task, *International Conference on Computer Communications and Networks*, pp.1-5, 2011.

[5] C. Huang, G. Chen, H. Yu, Y. Bao and L. Zhao, Speech emotion recognition under white noise, *Archives of Acoustics*, vol.38, no.4, pp.457-463, 2013.

[6] X. Huynh, T. Tran and Y. Kim, Convolutional neural network models for facial expression recognition using BU-3DFE database, *Information Science and Applications*, pp.441-450, 2016.

[7] S. Cheng, M. Chen, H. Chang and T. Chou, Semantic-based facial expression recognition using analytical hierarchy process, *Expert Systems with Applications*, vol.33, no.1, pp.86-95, 2007.

[8] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic and K. Scherer, Meta-analysis of the first facial expression recognition challenge, *IEEE Trans. Systems, Man, and Cybernetics, Part B: Cybernetics*, vol.42, no.4, pp.966-979, 2012.

[9] B. Efraty, C. Huang, S. K. Shah and I. A. Kakadiaris, Facial landmark detection in uncontrolled conditions, *International Joint Conference on Biometrics*, pp.1-8, 2011.

[10] T. Cootes, C. Taylor, D. H. Cooper and J. Graham, Active shape models – Their training and application, *Computer Vision and Image Understanding*, vol.61, no.1, pp.38-59, 1995.

[11] C. Santiago, J. Nascimento and J. Marques, 2D segmentation using a robust active shape model with the EM algorithm, *IEEE Trans. Image Processing*, vol.24, no.8, pp.2592-2601, 2015.

[12] B. van Ginneken, A. F. Frangi, J. J. Staal, B. M. Romeny and M. Viergever, Active shape model segmentation with optimal features, *IEEE Trans. Medical Imaging*, vol.21, no.8, pp.924-933, 2002.

[13] Z. Guo, L. Zhang and D. Zhang, A completed modeling of local binary pattern operator for texture classification, *IEEE Trans. Image Processing*, vol.19, no.6, pp.1657-1663, 2010.

[14] G. E. Hinton, S. Osindero and Y. W. Teh, A fast learning algorithm for deep belief nets, *Neural Computation*, vol.18, no.7, pp.1527-1554, 2006.

[15] M. J. Lyons, S. Akamatsu, M. Kamachi, J. Gyoba and J. Budynek, *The Japanese Female Facial Expression (JAFFE) Database*, http://www.kasrl.org/jaffe.html, 1998.