Action Unit Based Facial Expression Recognition Using Deep Learning

Salah Al-Darraji[1], Karsten Berns[1], and Aleksandar Rodić[2]

[1] Robotics Research Lab, Department of Computer Science, University of Kaiserslautern, Kaiserslautern, Germany.
{saleh,berns}@cs.uni-kl.de
[2] Robotics Laboratory, Mihailo Pupin Institute, University of Belgrade, Belgrade, Serbia.
aleksandar.rodic@pupin.rs

Abstract. Social interactive robot needs the same behaviors and capabilities of human to be able to work in human daily life. Humans usually use different types of verbal and nonverbal cues in their communication. Facial expressions are good examples of nonverbal cues used in inter-human interaction. This paper presents a facial expression recognition approach using deep learning. The approach is based on analysis of subtle changes in facial features of human face. The detected facial features, action units, are mapped to two psychological measurements, arousal and valence using support vector regression. Facial expression is then recognized by using these two values. The proposed approach has shown a recognition rate of more than 90%.

Keywords: Humanoid Robots, Facial Expression Recognition, Human-Robot Interaction, Deep Learning

1 Introduction

Facial expressions of humans play an important role in inter-human interaction. Usually, humans express more feelings through facial gestures than any other body movements. The emotional state of human can also be reflected on the face through facial expressions. Psychologists addressed diverse communicative functions for facial expressions such as giving feedback, opening and closing communication channels, and complementing verbal cues.

The facial action coding system (FACS) is the most widely used expression coding system in the behavioral sciences developed by Ekman and Friesen [1]. The basic element in this coding system is Action Unit (AU). Each action unit is related to the contraction of one or more facial muscles. FACS consists of 44 AUs which can occur in combinations as well. There are 12 AUs for upper face and 18 AUs for lower case. When more than one action unit are active together then a combination of action units occurs. More than 7000 combinations have been observed which can describe more complex facial actions. Action unit combinations can be either *additive* or *nonadditive*. Additive combination means that

the characteristics appearance of each AU in the combination is clearly recognizable and virtually unchanged. In nonadditive combinations, the characteristics appearance of all AU involved in the combination is changed and cannot be recognized separately.

The facial expression recognition approaches based on AUs differ in either the feature extraction techniques or the classification technique, or both. Valstar and Pantic [9], for example, detect only 15 action units that are relevant to the six basic emotions. They use facial feature points detector based on Gabor wavelets and GentleBoost classifiers to localize 20 fiducial facial points in the first frame and then tracked in all subsequent frames. Then the calculated features are given to 15 SVMs that are trained using features that describe the spatio-temporal relationships between the tracked points.

Tong et. al. [8] focus on the temporal evolution and the semantic relationships among action units. They use Dynamic Bayesian Network (DBN) to model the relationships among different AUs. They use a set of multi-scale and multi-orientation Gabor wavelets to calculate wavelet features for each pixel. Then AdaBoost classifier combines the wavelet features to produce a measurement score for each AU. The drawback of this work is the high complexity and therefore, cannot be used in real time systems.

Khademi et. al. [2] recognize single AUs and AU combinations using both geometric and appearance features. To detect the geometrical features, they place manually a point grid on the first frame and register this grid automatically with the face. They use optical flow tracker to track these points in the successive frames and calculate the displacement with respect to the first frame. They extract the appearance facial features by using a set of Gabor wavelet and then use a mixture of HMMs and neural networks to recognize subtle changes in AUs. Although this method is robust, however, the major drawback is the requirement to manually annotate the first frame.

This paper proposes an action unit based facial expression recognition using Convolutional Neural Network (CNN). Using 23 deep classifiers, the most relevant facial action units and combinations are detected. To reduce the dimensionality, two Support Vector Machines for Regression (SVRs) are used to map these action unit activations to only two values, arousal and valence. These values of arousal and valence are then used to determine the facial expression according to psychological studies. A genetic algorithm is used to select the proper parameters set for the CNN.

The rest of the paper is organized as follows: Section. 2 explains the proposed approach. Section. 3 discusses experiments and result. Finally, Sect. 4 provides the conclusion.

2 Facial expression recognition

Most of the existing facial expression analysis approaches focus on recognizing a small set of prototypic facial expressions, i.e. happiness, sadness, fear, anger, surprise, and disgust. In inter-personal interaction, many nonverbal cues that

occurs due to subtle changes in the facial features, can not be regarded as one of these prototypic expressions. In order to detect and interpret the small changes in the facial features, recognition of facial features (AUs) is needed.

2.1 Action units analysis

In this work, we use deep neural network, specifically convolutional neural network (CNN), to recognize each of the related action units and combinations. CNN consists of multiple different types of layers that are connected to each other. Each layer can be either convolution layer, pooling layer, or fully connected layer.

The convolutional layer applies a set of learnable filters on the input image. Using more than one convolutional layer enables to extract features on different levels. For $m \times m$ filter w, the output unit x_{ij}^{l} of the convolutional layer l is calculated by using 1.

$$x_{ij}^{l} = \sum_{a=0}^{m-1} \sum_{b=0}^{m-1} w_{ab} y_{(i+a)(j+b)}^{l-1} \ . \tag{1}$$

After calculating the convolution, nonlinearity is applied to calculate y_{ij}^{l} using the activation function.In this paper, we have used hyperbolic tangent ($tanh$) function in all layers.

Usually, after each convolution layer a pooling layer is used for down-sampling the input image nonlinearly. We have used max pooling, in which a $k \times k$ region is reduced to a single value, which is the maximum value in that region. For an input layer of $N \times N$, the output of max pooling layer is $\frac{N}{k} \times \frac{N}{k}$.

After a set of pairs convolution and pooling layers, the high level reasoning is done by one or more fully connected layer. The fully connected layers represent a multilayer perceptron, which works as classifier. No spatial information exists after fully connected layer and it can be represented as one dimensional layer.

Fig. 1 depicts the architecture of the convolutional neural network that is used to recognize facial action units. The architecture comprises 6 layers (excluding input layers). It receives a human face as a 32×32 gray image and outputs the confidences of seven facial expressions (including neutral). It uses two convolutional layers each with its own sub-sampling layer (max-pooling).

In learning phase, mean squared error (MSE) is used as a loss function. The optimization function used in this work is the Levenberg-Marquardt gradient descent, which combines the advantages of the steepest descent method with the Gauss-Neuton method by adaptively varying the parameter updates between the two methods.

2.2 Arousal-valence estimation

Recognizing action units as basic components is very important to recognize various nonverbal cues such as facial expressions and some other nonverbal cues. Paltoglou [5] mapped a subset of emotions to a 2D Cartesian coordinate system

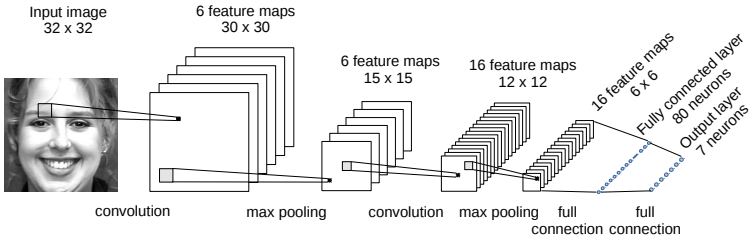


Fig. 1: Architecture of deep neural network for action unit recognition. It contains two convolutional layers each with its maximum sampling layer and two fully connected layers with 80 and 7 neurons respectively.

according to the circumplex model of Russell [6]. In this mapping, emotion states are represented as points where x-axis is *valence* and y-axis is *arousal*. Using this representation in computer vision systems enables to detect emotions with their corresponding intensity. It also enables to detect different emotions and is not restricted to some prototypic ones.

The challenge of using the arousal-valence representation is calculating these two values. In this work, the values are calculated using AU values of human face. Two SVRs, specifically ϵ-SVRs, are used to calculate arousal and valence values. The input of these two SVRs are the 20 action unit values in addition to 3 action unit combinations: 1+4, 1+2+4, and 4+5. The aim of ϵ-SVR is to find a function $f(x)$ that has at most ϵ deviation from the targets y_i for all training data and is as *flat* as possible [7].

On a general scale and not limited to the prototypic ones, activation of AUs of human face is used to recognize the facial expression. Suppose, U is a set of n action unit activations that are involved in facial expressions, where

$$U = \{u_1, u_2, \ldots, u_n\}, \quad with \quad u_i \in [0, 1], n > 0 \ . \tag{2}$$

P is a set of deep learners (CNNs) that detects the activation of all AUs u_i of an input image x .

$$P = \{p_1, p_2, \ldots, p_n\} \ . \tag{3}$$

$$u_i = p_i(x) \quad where \quad x \in \mathbb{R} \times \mathbb{R} \ . \tag{4}$$

Suppose, A and V are two SVRs to calculate the *arousal* and *valence* respectively. Then a and v are the arousal and valence values of input image x.

$$a = A(u_1, u_2, \ldots, u_n) \ , \qquad v = V(u_1, u_2, \ldots, u_n) \ . \tag{5}$$

The function f that calculates the facial expression is defined as follows:

$$f(a, v) = \arg\min_{i \in E} \left(\sqrt{(a - a_i')^2 + (v - v_i')^2}\right) \ . \tag{6}$$

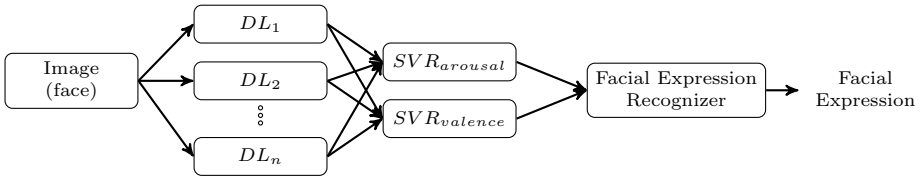


Fig. 2: Facial expression recognition process. DL_i refers to the deep learner for action unit i. $SVR_{arousal}$ and $SVR_{valence}$ are the SVRs for arousal and valence respectively.

Where E is all facial expressions, a'_i and v'_i are arousal and valence of expression E_i. Fig. 2 shows the facial expression recognition process. The training process of deep learners and SVRs are discussed in Sect. 3.

3 Experiments

3.1 Datasets

To evaluate the performance of the proposed facial expression recognition approach, two faces datasets are used: CK+ (Cohn-Kanade plus) [4] and RaFD (Radboud Faces Database) [3]. CK+ provides 593 sequences of frontal faces in still images from 123 subjects each with action units annotation. In our study, CK+ dataset is used to train the AU classifiers. For each action unit, a deep learner has been trained to classify two classes: *active* and *not active*. The deep learners output the confidence value of active c_a and not active c_n classes, where $c_a, c_n \in [0, 1]$. The activation value of each action unit refers to how much this AU is active, which can be calculated as follows:

$$activation = c_a \cdot (1 - c_n) \ . \tag{7}$$

RaFD (Radboud Faces Database) is a high quality faces dataset that contains images of 67 subjects displaying 8 facial expressions (neutral, anger, disgust, fear, happiness, sadness, surprise, and contempt). It also provides additional validation data such as agreement percentage, intensity rating, clarity, genuineness, and valence rating for each image. It doesn't provide arousal value of images. In order to calculate arousal value of each image, the standard arousal of the corresponding facial expressions can be weighted using the intensity rating as described by [5].

3.2 Parameters optimization for CNNs

Many parameters play an essential role in the design of any deep neural network. Number of layers, number of neurons in each layer, activation function, and cost function are examples of these parameters. Setting these parameters

affects the learning process and the performance of the network. In this work, a Genetic Algorithm (GA) based optimization approach is used to set all network parameters.

Genetic algorithm is a heuristic search algorithm that mimics the process of natural selection. GA repeatedly modifies a population of individual solutions (chromosomes). At each step, GA selects parent chromosomes from the current population to produce the children for the next generation. The optimization problem in the current work is a maximization of the recognition rate of facial action units. The main ingredients of GA are as follows:

Chromosome In genetic algorithms, chromosome is a set of genes, which codes the independent variables as a solution of the given problem. In this optimization problem, a chromosome is a set of parameters that represents the structure of the network (classifier) such as *number of layers*, *kernel size*, *pooling size*, *activation function*, and so on. The size of chromosome is variable and depends on the number of convolutional and fully connected layers.

Selection This process allows for better individual to contribute to the next generation. The goodness of an individual depends on its fitness. This can be determined by a fitness function, which is the recognition rate in this work.

Crossover This process combines two individuals as parents to form children for the next generation. In this work, one point is randomly chosen, the values after this point are exchange to produce two children. Because of the different size of each chromosome, only the genes of one chromosome that have corresponding genes in the other one is exchanged.

Mutation This means random change in the value of one or more genes with very low probability. It can maintain diversity within population and inhibit premature convergence.

All the above GA operations have been applied on a population of 50 chromosomes to obtain an optimum solution for the facial action units recognition. After 100 generations, the fittest chromosome is the network configuration that has been shown in Fig. 1.

3.3 Evaluation results

To evaluate the recognition of action units, 23 classifiers have been trained to recognize two classes for each action units: *active* and *not active*. We have used leave-one-subject-out cross-validation to maximize the training and testing data. The overall recognition accuracy is more than 90%. Table 1 shows the recognition rate of action units.

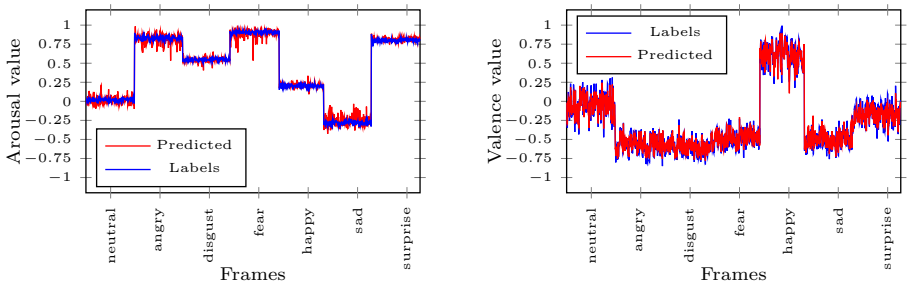
In the same way, leave-one-subject-out cross-validation configuration is used to evaluate the two SVRs. The root mean square for the *arousal* and *valence*

Table 1: Action units recognition rates. The average is weighted i.e. depending on the number of positive examples.

AUs	Samples	RR (%)	AUs	Samples	RR (%)	AUs	Samples	RR (%)
1	88	94.32	11	39	84.62	24	65	80.77
2	90	94.44	12	126	96.03	25	317	93.38
4	103	91.26	15	93	89.78	26	50	80.00
5	86	92.44	16	24	81.25	27	80	100.00
6	92	96.20	17	172	88.37	1+4	67	85.07
7	144	89.58	20	73	92.47	1+2+4	21	71.43
9	71	95.07	22	14	89.29	4+5	25	74.00
10	40	92.50	23	65	85.38			
Average		**90.85**						

are 0.21 and 0.16, respectively. Fig. 3 shows arousal values (a) and the valence values (b) of all examples categorized by the facial expression.

The performance of the whole facial expression recognition using the proposed approach is given in Table 2. It is obvious from this table that *angry* and *fear* are overlapped because they are near to each other on the arousal-valence map. An average recognition rate of 90.3% is achieved using this approach.



(a) Arousal. (b) Valence.

Fig. 3: Arousal and valence regression.

4 Conclusion

In this paper, we have presented a facial recognition approach based on AUs. The most relevant 23 AUs and combination have been used in this work. This approach uses deep learning in action unit analysis. For each of these action units and combination, a two-class classifier has been trained. Two SVRs have been

Table 2: Confusion matrix of facial expression recognition.

Detected as ⇒	Neutral	Angry	Disgust	Fear	Happy	Sad	Surprise
Neutral	95.2	0.0	0.0	4.8	0.0	0.0	0.0
Angry	0.0	81.0	9.5	4.8	0.0	4.8	0.0
Disgust	0.0	0.0	95.2	4.8	0.0	0.0	0.0
Fear	0.0	9.5	0.0	81.0	0.0	0.0	9.5
Happy	0.0	0.0	0.0	0.0	94.1	0.0	5.9
Sad	4.8	0.0	4.8	0.0	0.0	90.5	0.0
Surprise	0.0	0.0	0.0	4.8	0.0	0.0	95.2

used to calculate the values of arousal and valence of the human depending on the activation values of these action units and combinations. The facial expression is recognized by finding the minimum Euclidean distance from the detected point (arousal and valence) and the basic emotions. The recognition rate of both AUs and facial expression is more than 90%.

References

1. Ekman, P., Friesen, W., Hager, J.: Facial Action Coding System. Consulting psychologist Press, Inc, (1978)
2. Khademi, M., Manzuri-Shalmani, M. T., Kiapour, M. H., Kiaei, A. A.: Recognizing combinations of facial action units with different intensity using a mixture of hidden markov models and neural network. In: Proceedings of the 9th International Conference on Multiple Classifier Systems, MCS'10, pp. 304–313, (2010)
3. Langner, O., Dotsch, R., Bijlstra, G., Wigboldus, D. H. J., Hawk, S. T., van Knippenberg, A.: Presentation and validation of the radboud faces database. Cognition and Emotion, 24(8):1377–1388, (2010)
4. Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar,Z., Matthews, I.: The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), pp. 94–101, (2010)
5. Paltoglou, G., Thelwall, M.: Seeing stars of valence and arousal in blog posts. IEEE Transactions on Affective Computing, 4(1):116–123, (2013)
6. Russell, J.: The circumplex model of affect. Journal of Personality and Social Psychology, 39:1161–1178, (1980)
7. Smola, A. J., Schölkopf, B.: A tutorial on support vector regression. Statistics and Computing, 14(3):199–222, (2004)
8. Tong, Y., Liao, W., Ji, Q.: Facial action unit recognition by exploiting their dynamic and semantic relationships. Pattern Analysis and Machine Intelligence, IEEE Transactions, 29(10):1683–1699, (2007)
9. Valstar, M., Pantic M.: Fully automatic facial action unit detection and temporal analysis. In: Computer Vision and Pattern Recognition Workshop (CVPRW), pp. 149–149. (2006)