International Journal of Advanced Technology & Engineering Research (IJATER)
www.ijater.com

# FACIAL EXPRESSION RECOGNITION BASED ON DEEP EVOLUTIONAL SPATIAL-TEMPORAL NETWORKS

Gowtam Reddy Gurrala[1], Ajay Duddu[2], Shruti Bhargava Choubey[3]
[1,2] Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, 501301, Telangana, India
[3]Associate Professor, Department of Electronics and Communication Engineering, Sreenidhi Institute of Science and Technology, 501301, Telangana, India
[1]gowtamreddy543@gmail.com
[2]aaj52539@gmail.com
[3]shrutibhargava@sreenidhi.edu.in

***Abstract*—One key challenging issue of facial expression recog- nition is to capture the dynamic variation of facial physical structure from videos. In this paper, we propose a Part-based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) to analyze the facial expression information of temporal se- quences. Our PHRNN models facial morphological variations and dynamical evolution of expressions, which is effective to extract "temporal features" based on facial landmarks (geometry information) from consecutive frames. Meanwhile, in order to complement the still appearance information, a Multi-Signal Convolutional Neural Network (MSCNN) is proposed to extract "spatial features" from still frames. We use both recognition and verification signals as supervision to calculate different loss functions, which are helpful to increase the variations of different expressions and reduce the differences among identical expressions. This deep Evolutional Spatial-Temporal Networks (composed of PHRNN and MSCNN) extract the partial-whole, geometry-appearance and dynamic-still information, effectively boosting the performance of facial expression recognition. Experimental results show that this method largely outperforms the state-of-the-art ones. On three widely used facial expression databases (CK+, Oulu-CASIA and MMI), our method reduces the error rates of the previous best ones by 45.5%, 25.8% and 24.4%, respectively.**

***Index Terms*—Facial expression recognition, dynamical evolution, recognition and verification signals, deep Spatial-Temporal Networks**

## I. INTRODUCTION

Emotions often mediate and facilitate interactions among human beings. Thus, understanding emotion often brings context to seemingly bizarre and/or complex social communication. Emotion can be recognized through a variety of means such as voice intonation, body language, and more complex methods such electroencephalography (EEG). However, the easier, more practical method is to examine facial expressions. There are seven types of human emotions shown to be universally recognizable across different cultures: anger, disgust, fear, happiness, sadness, surprise, contempt. Interestingly, even for complex expressions where a mixture of emotions could be used as descriptors, cross- cultural agreement is still observed. Therefore, a utility that detects emotion from facial expressions would be widely applicable. Such an advancement could bring applications in medicine, marketing and entertainment.

The task of emotion recognition is particularly difficult for two reasons:

1) There does not exist a large database of training images and

2) classifying emotion can be difficult depending on whether the input image is static or a transition frame into a facial expression.

The latter issue is particularly difficult for real-time detection where facial expressions vary dynamically.

Early researches about facial expression recognition mainly focus on recognizing expressions from still frames. These methods effectively extract spatial information but cannot model the variability in morphological and contextual factors. Recently, some studies try to capture the dynamic variation of facial physical structure from consecutive frames based on hand-crafted features or deep learning methods, such as LBP-TOP, HOG 3D, STM-ExpLet, and DTAGN.

In this paper, we propose a Multi-Signal Convolutional Neural Network (MSCNN) to extract spatial features from still frames. Instead of only using a recognition signal as supervision, our MSCNN is trained under the supervision of recognition and verification signals. The two signals corresponding to different loss functions are helpful to increase the variations of different expressions and reduce the difference among identical

expressions, which can force our model to focus on expression itself regardless of different subjects, illuminations, ages and so on. Due to the MSCNN model, we can capture the whole, appearance and still information. Finally, we fuse the MSCNN and PHRNN to the Evolutional Spatial-Temporal Networks to make the final decision.

## II. Facial Expression Evolutionary Reasons

A common assumption is that facial expressions initially served a functional role and not a communicative one. Here, each one of the seven classical expressions with its functional initially role are justified:

**Anger:**

It involves three main features- teeth revealing, eyebrows down and inner side tightening, squinting eyes. The function is clear- preparing for attack. The teeth are ready to bite and threaten enemies, eyes and eyebrows squinting to protect the eyes, but not closing entirely in order to see the enemy.

Fig1: Anger

**Disgust:**

It involves wrinkled nose and mouth. Sometimes even involves tongue coming out. This expression mimics a person that tasted bad food and wants to spit it out, or smelling foul smell.

Fig2: Disgust

**Fear:**

It involves widened eyes and sometimes open mouth. The function- opening the eyes so wide is suppose to help increasing the visual field (though studies show that it doesn't actually do so) and the fast eye movement, which can assist finding threats. Opening the mouth enables to breath quietly and by that not being revealed by the enemy.

Fig3: Fear

**Surprise:**

It is very similar to the expression of fear. Maybe because a surprising situation can frighten us for a brief moment, and then it depends whether the surprise is a good or a bad one. Therefore, the function is similar.

Fig4: Surprise

**Sadness:**

It involves a slight pulling down of lip corners, inner side of eyebrows is rising. Darwin explained this expression by suppressing the will to cry. The control over the upper lip is greater than the control over the lower lip, and so the lower lip drops. When a person screams during a cry, the eyes are closed in order to protect them from blood pressure that accumulates in the face. So, when we have the urge to cry and we want to stop it, the eyebrows are rising to prevent the eyes from closing.

Fig5: Sadness

**Contempt:**

It involves lip corner to rise only on one side of the face. Sometimes only one eyebrow rises. This expression might look like half surprise, half happiness. This can imply the person who receives this look that we are surprised by what he said or did (not in a good way) and that we are amused by it. This is obviously an offensive expression that leaves the impression that a person is superior to another person.

Fig6: Contempt

**Happiness:**

It usually involves a smile- both corner of the mouth rising, the eyes are squinting and wrinkles appear at eyes corners. The initial functional role of the smile, which represents happiness, remains a mystery. Some biologists believe that smile was initially a sign of fear. Monkeys and apes clenched teeth in order to show predators that they are harmless. A smile encourages the brain to release endorphins that assist lessening pain and resemble a feeling of well being. Those good feeling that one smile can produce can help dealing with the fear. A smile can also produce positive feelings for someone who is witness to the smile, and might even get him to smile too. Newborn babies have been observed to smile involuntarily, or without any external stimuli while they are sleeping. A baby's smile helps his parents to connect with him and get attached to him. It makes sense that for evolutionary reasons, an involuntary smile of a baby helps creating positive feelings for the parents, so they wouldn't abandon their offspring.

Fig7: Happiness

## III. Related work

The task of recognizing facial expressions is usually subdivided into three subordinate challenges: face detection, feature extraction, and expression classification [3]. The first step aims at determining the position and shape of the face in the image. Features descriptive for facial expressions or head gestures are extracted in the second step. In the third step a classifier is applied to the features to identify the expression class.

**Facial expression recognition:**

Facial expression recognition methods can be classified in two categories: frame-based and sequence-based methods. Earlier research mostly focuses on expression analysis based on still frames. However, these methods are unable to successfully model the variability in morphological and contextual factors. As a dynamic event, recognizing facial expression from consecutive frames is more natural and proved to be more effective in recent years. Traditional hand-crafted features are extended to adapt to consecutive frames, such as 3D- HOG, LBP-TOP, 3D-SIFT. Among all the traditional methods, Guo et al. propose a longitudinal atlases construction which achieves the best performance on the Oulu-CASIA database. In order to extract more powerful spatio-temporal features, Liu et al. propose an expressionlet-based spatio-temporal mainfold descriptor which outperforms the previous traditional methods on the CK+ and MMI databases. The three databases are widely used and most sequences in them

contain more than 10 frames to reflect the gradual variation of expression. Therefore, we also choose to do experiments on them rather than other databases.

**Deep Neural Networks:**

Recurrent Neural Network has many successful applications for modeling of sequential data such as handwriting recognition, gesture recognition and video description. Several researchers attempt to utilize RNN to solve the problem of expression recognition. They input the facial images into RNN directly to capture the dynamic variations of facial structure. Meanwhile, Jung et al. utilize a small DNN to capture the dynamical variations of expression. Their proposed DTAGN method achieves the best performance on the CK+ and Oulu-CASIA databases, even exceeds all hand-crafted methods. While the encouraging results are obtained, there are still shortcomings among these works: 1) It is hard for neural networks to model the dynamical variations of expression without any prior knowledge and constraints, especially on a small database.

2) A small deep model cannot take full advantage of the deep learning methods to extract high-level temporal features. To address these problems, here we propose a deep PHRNN to capture the temporal information by modeling the facial morphological variations and dynamically evolutional properties of expression.

Over the past few years, models based on deep convolutional network have dominated various vision tasks, such as image classification, objection recognition and face analysis. For the task of facial expression recognition, a relevant study is 3DCNN-DAP. A deformable parts learning component is incorporated into the 3DCNN framework to capture the expression features from motion. Similar to 3DCNN-DAP, Jung et al. propose a small CNN to capture the dynamical variations of appearance. While CNN has achieved reasonably good performance in expression recognition, there are two shortcomings among these methods:

1) The recognition signal can pull apart the features of different expressions since they have to be classified into different classes, but it has not a strong constraint to reduce the same-expression variations.

2) Faces of the same expression have much difference when they are presented by different people under different illuminations, ages and so on.

It is hard to push the deep model to focus on the expression itself on small databases. To address these problems, we propose a MSCNN to learn the spatial features by using both recognition and verification signals as supervision. The signals correspond to different loss functions, which are beneficial to force the model to focus on expression information for learning powerful features.

## IV. Proposed model

The proposed Spatial-Temporal Networks include two kinds of networks. Firstly, we extend PHRNN to a temporal network to capture dynamic features from consecutive frames. Secondly, a spatial network based on MSCNN is constructed to extract static features from still frames. Finally, the two kinds of networks are combined to improve the performance of facial expression recognition

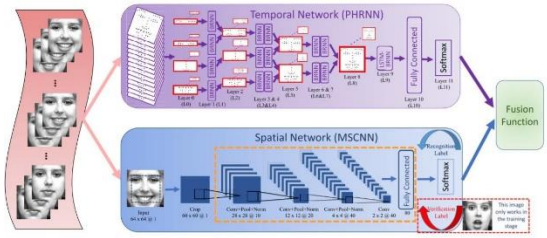


Fig8: Our proposed Spatial-Temporal Networks for facial expression recognition.

Temporal network (PHRNN): facial landmarks are divided into four parts based on facial physical structure, and then separately fed into our model. Local features are concatenated along the feature extraction cascade, while the global high-level features are formed in the upper layers according to facial morphological variations and dynamically evolutional properties. Spatial network (MSCNN): in the training stage, our MSCNN takes pairs of frames as the input with both recognition and verification signals as supervision, which is helpful to increase the variations of different expressions and reduce the difference of identical expressions. The two signals correspond to different loss functions which help to force our model to focus on expression itself, rather than other factors such as identities and illuminations.

## V. Part based Hierarchical Bi-Directional Recurrent Neural Network

In this section, we describe our proposed PHRNN model. Traditional RNN learns complex temporal dynamics by mapping an input sequence x to a sequence of hidden states. The hidden states of a recurrent layer h and the output of a single hidden layer RNN z can be expressed as:

$$ht = H(Wxhxt + Whh\ xt-1 + bh)$$

$$zt = O(Whzht + bz)$$

where Wxh, Whh, Whz are the connection weights from the input layer to the hidden layer, bh and bz are two biases of the hidden layer and the output layer. H(·) and O(·) are the activation functions. One shortcoming of conventional RNN is that it is difficult to learn long-term dynamics due to the vanishing gradient problem. Long-Short Term Memory, contains self-connected memory units, and provides a solution to explore long range contextual information of complex temporal dynamics [38]. The activation of the memory cell is implemented by the following composite functions:

$$it = \sigma\ (Wxi\ xt + Whi\ ht-1 + Wci\ ct-1 + bi)$$

$$ft = \sigma\ (Wx\ f\ xt + Wh\ f\ ht-1 + Wc\ f\ ct-1 + bf)$$

$$ct = ft\ ct-1 + it\ \tanh(Wxcxt + Whcht-1 + bc)$$

$$ot = \sigma\ (Wxoxt + Whoht-1 + Wcoct + bo)$$

$$ht = ot\ \tanh(ct)$$

where σ (·) is the sigmoid function, i, f, and o denote the input gate, forget gate and output gate, respectively. All of the matrices W are the weights between two gates.

Usually, it only utilizes the past context in the sequence. In facial expression recognition, future context should also be taken into account. Schuster and Paliwal propose the bidirectional recurrent neural network (BRNN), which can process data in both directions and then fed them into the same output layer. Benefiting from the power of BRNN to store and access to the long-range contextual information, we propose a Part based Hierarchical Bidirectional Recurrent Neural Network (PHRNN) for facial expression recognition. The temporal information is the variations of the facial critical areas implied in sequential frames, which can be well mapped to facial landmarks. According to facial physical structure of a human face, we divide facial landmarks into four parts, i.e., eyebrows, eyes, nose and mouth. All of facial expressions can be performed by these parts. For example, happiness causes the corners of the lips up, disgust causes the eyebrows and eyes shrink, while surprise can be decomposed to eyes larger with mouth open widely. In order to learn powerful features from the facial critical areas, the four parts of landmarks are fed into four BRNN subnets, respectively.

## VI. Multi Signal Convolutional Neural Networks

Our architecture contains four convolutional layers, a fully-connected layer and a softmax layer. The input is 60×60 gray images. Following the input, the first convolutional layer is generated after convolving the input via 10 filters of a size 5×5×1 with a stride of 1 pixel. The second convolutional layer filters the output of the previous layer with 20 kernels of a size 5×5×10 and the third convolutional layer contains 40 kernels of a size 5×5×20, both with a stride of 1 pixel. Each of the first three convolutional layers is followed by a max-pooling layer and a local response normalization layer, which is helpful to increase the translation invariance and avoid overfitting. The fourth stage contains only a convolutional layer using 80 filters of a size 3×3×40. Finally, the expression descriptor is extracted by a fully-connected layer with 80 neurons, and fed into a softmax layer to classify.

In term of the training time with the gradient descent algorithm, the non-saturating nonlinearity f (x) = max(0, x) is much faster than the saturating nonlinearity . Thus, we adopt the ReLU function as the activation function of neurons, which has achieved better performance than the sigmoid function.

The two signals in the MSCNN corresponds to two loss functions which work together to push our model to focus on expression information, rather than identities, illuminations, ages and son on.

## VII. Conclusion

In this paper, Evolutional Spatial Temporal Networks to extract multiple kinds of features for facial expression recognition is proposed. Specially, according to

the facial morphological variations and dynamically evolutional properties, we presented PHRNN to capture the dynamic variation of facial physical structure from videos. In order to complement the static appearance information, we propose MSCNN with two signals to increase the variations of different expressions and reduce the differences among identical expressions. The two kinds of networks capture the partial-whole, geometry appearance and dynamic-still information simultaneously, and complement each other to boost the performance of recognition. Experimental results on three databases demonstrate that our proposed methods have achieved the state-of-the-art performance.

*REFERENCES:*

[1]. Sudarshan Adeppa, ―Detection of Objects across the Walls with Wi-Fi Technology‖, International Journal on Emerging Technologies, 2015.

[2]. Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in Proc. Comput. Vis. (ECCV), 2012, pp. 631–644.

[3].R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR), Jun. 2014, pp. 580–587

[4].K. SIMONYAN AND A. ZISSERMAN, "VERY DEEP CONVOLUTIONAL NETWORKS FOR LARGE-SCALE IMAGE RECOGNITION," IN PROC. INT. CONF. LEARN. REPRESENT. (ICLR), APR. 2015