This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2689999, IEEE Transactions on Image Processing

Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks

Kaihao Zhang, Yongzhen Huang, *Member, IEEE,* Yong Du, *Student Member, IEEE,* and Liang Wang, *Senior Member, IEEE*

***Abstract*—One key challenging issue of facial expression recognition 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 sequences. 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

FACIAL expression recognition has become an increasingly active research topic in the field of computer vision, as it plays an important role in many applications such as human-computer interaction [1] and health care [2]. Early researches about facial expression recognition mainly focus on recognizing expressions from still frames [3]. 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 [4], HOG 3D [5], STM-ExpLet [6], and DTAGN [7].

However, as a special facial analysis task, facial expression recognition has its own characteristic. In particular, facial expression can be considered as dynamic variation of key parts (*e.g.* eyes, nose and mouth), which are fused to form the variation of local parts and the whole face, and the key challenge becomes to capture such dynamic variation of facial physical structure from consecutive frame. For traditional methods, it is hard to extract powerful temporal features hidden in facial images based on hand-crafted descriptors. Also, inputting the facial images into deep neural networks directly is unable to effectively utilize the prior knowledge as the above mentioned. This makes against to learn the evolutional properties of an expression.

In this paper, considering that the variation of facial landmarks is an important representation of facial expressions and deep RNN [8], [9], [10] has the advantage of modelling the contextual information of temporal sequences, we propose a Part-based Hierarchical Recurrent Neural Network (PHRNN) to extract temporal features based on facial landmarks from motion over time. Our proposed PHRNN models the facial morphological variations and the evolutional properties of expression, which is effective to capture the dynamic variation of the facial physical structure. The facial landmarks are divided into four parts based on the facial structure. Each subnet takes a part of landmarks as the input and extracts the partial low-level features in the bottom layers. According to the evolutional properties of expression, these features continue to concatenate along the feature extraction cascade while the local and global high-level features are formed in the upper layers. Finally, the temporal features of a facial expression are obtained. Due to the PHRNN model, we can capture the partial, geometry and temporal information.

As a special video classification task, the still frame of a facial expression video has strong discrimination. CNNs [11], [12], [13], [14], [15] have been used as an effective model for image recognition tasks. However, a small database is often a large impediment to apply these CNN models directly to recognize facial expression. Proposing deep neural networks with a small number of hidden layers is a feasible way to overcome the over-fitting problem [7], but it does not benefit to take advantage of the deep learning methods to extract deep high-level features. Meanwhile, how to force the neural networks to focus on expression information rather than identities or other factors is also a question.

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 corre-

K. Zhang is with the College of Engineering and Computer Science, the Australian National University, Canberra, ACT, Australia. Y. Huang, Y. Du and L. Wang are with National Laboratory of Pattern Recognition (NLPR), Center for Excellence in Brain Science and Intelligence Technology (CEBSIT), Institute of Automation, Chinese Academy of Sciences (CASIA), University of Chinese Academy of Sciences (UCAS), Beijing, 100190, China. Y. Huang is the corresponding author. E-mail: dr.khzhang@gmail.com {yong.du, yzhuang, wangliang}@nlpr.ia.ac.cn.

sponding 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.

The main contributions of this paper are three-fold. Firstly, we propose a PHRNN model to extract dynamic geometry information. Landmarks are decomposed into different parts according to the facial morphological variations, which are helpful to model dynamical evolution of expression. Secondly, in order to complement the still appearance information, we propose a MSCNN model with both recognition and verification signals used as supervision. The two signals correspond to two different loss functions, which are helpful to increase the variations of different expressions and reduce the difference among identical expressions. Thirdly, the PHRNN and MSCNN complement each other to compose the Evolutional Spatial-Temporal Networks, which considers partial-whole, geometry-appearance and dynamic-still information simultaneously. Experimental results demonstrate that our proposed method outperforms the previous best methods in facial expression recognition, with a large improvement on three widely used facial expression databases, *i.e.*, 45.5% on CK+, 25.8% on Oulu-CASIA and 24.4% on MMI, respectively.

II. Related Work

A. Facial Expression Recognition

Facial expression recognition methods can be classified in two categories: frame-based and sequence-based methods [16]. Earlier research mostly focuses on expression analysis based on still frames [17], [18], [19]. 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 [6], [20], [21], [22]. Traditional hand-crafted features are extended to adapt to consecutive frames, such as 3D-HOG [5], LBP-TOP [20], 3D-SIFT [23]. Among all the traditional methods, Guo *et al.* [24] propose a longitudinal atlases construction which achieves the best performance on the Oulu-CASIA database [25]. In order to extract more powerful spatio-temporal features, Liu *et al.* [6] propose an expressionlet-based spatio-temporal mainfold descriptor which outperforms the previous traditional methods on the CK+ [26] and MMI databases [27]. 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.

B. Deep Neural Networks

Recurrent Neural Network has many successful applications for modeling of sequential data such as handwriting recognition [28], [29], gesture recognition [30] and video description [10]. Several researchers attempt to utilize RNN to solve the problem of expression recognition [31], [32], [33]. They input the facial images into RNN directly to capture the dynamic variations of facial structure. Meanwhile, Jung *et al.* [7] 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 [11], [12], [15], objection recognition [13] and face analysis [34], [35], [36]. For the task of facial expression recognition, a relevant study is 3DCNN-DAP [37]. 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.* [7] 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.

III. Our 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. Figure 1 shows the architecture of facial expression recognition used in this paper.

A. PHRNN for Modeling Dynamical Evolution

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

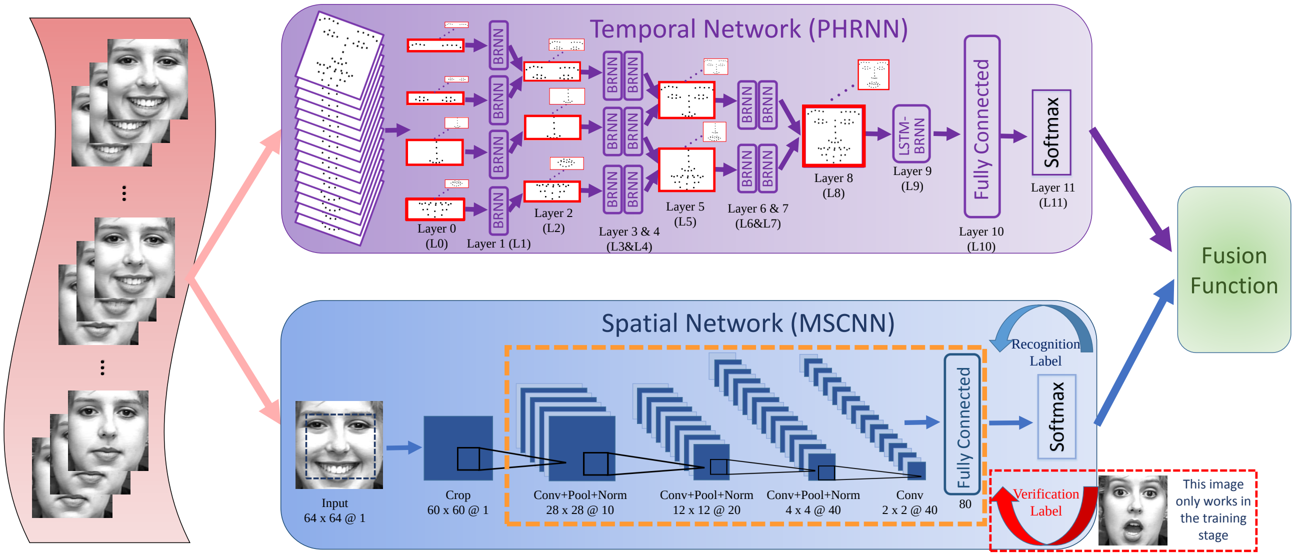


Fig. 1. Our proposed Spatial-Temporal Networks for facial expression recognition. Temporal network (PHRNN): facial landmarks are divided into four parts based on facial physical structure, and then separately fed into our model. Local features are concatenated along the feature extraction cascade, while 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. Please refer to Figure 2, Section III-B and Section IV-C to get more details about the training stage. In the testing stage, our proposed MSCNN takes one current frame as input.

hidden states of a recurrent layer h and the output of a single hidden layer RNN z can be expressed as:

$$h_t = H(W_{xh}x_t + W_{hh}x_{t-1} + b_h) \tag{1}$$

$$z_t = O(W_{hz}h_t + b_z) \tag{2}$$

where W_{xh}, W_{hh}, W_{hz} are the connection weights from the input layer to the hidden layer, b_h and b_z are two biases of the hidden layer and the output layer. $H(\cdot)$ and $O(\cdot)$ 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:

$$i_t = \sigma(W_{xi}x_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{3}$$

$$f_t = \sigma(W_{xf}x_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{4}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}x_t + W_{hc}h_{t-1} + b_c) \tag{5}$$

$$o_t = \sigma(W_{xo}x_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{6}$$

$$h_t = o_t tanh(c_t) \tag{7}$$

where $\sigma(\cdot)$ 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 *et al.* [39] 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.

Morphological variations and dynamically evolution. As the above mentioned, facial expression can be considered as dynamic variation of key parts, which are fused to form the variation of local parts and the whole face. Therefor one key challenge becomes to capture such dynamic variation of facial physical structure from consecutive frame. Our PHRNN is proposed based on this idea, which is shown in the upper part of Figure 1. Local features are concatenated along the feature extraction cascade, while the global high-level features are formed in the upper layers based on the facial morphological variations and dynamically evolutional properties of expression. Specifically, each subnet extracts one part of local low-level features in the L1. To model the neighboring landmarks, we combine the representations of eyebrows and eyes to obtain a new representation in the L2. Followed by two BRNNs in the L3 and L4, we obtain the features of the eyebrow-eye, nose and mouth. The representations of eyebrow-eye and nose are concatenated to obtain the upper half face while the representations of nose and mouth are concatenated to obtain the bottom half face in the L5, then fed into two BRNNs in the L6 and L7. We can obtain the representation of the whole face in the L8. The temporal dynamics of the whole face are fed into a BRNN in the L9 and a fully connected layer in the L10. Finally, the softmax layer is used to estimate the facial expression, taking the following form:

$$P(c|w_i) = \frac{e^{a_c(w_i)}}{\sum_{l=0}^{C-1} e^{a_l(w_i)}} \tag{8}$$

where w_i is obtained from the last hidden layer for the corresponding target c, and $a_c(w_i)$ is the accumulated result. l is one of facial expression classes.

For training, our goal is to minimize the maximum likelihood. The objective function of our model is the cross entropy error function:

$$\mathcal{L} = -\sum_{c=1}^{C} z_c ln P(c|w_i) \tag{9}$$

where $z_c \in \{0, 1\}$. $P(c|w_i)$ is the predicted probability of the facial expression c.

Implementation details and overfitting problem. Many parameters may affect the performance of PHRNN such as the number of neurons in each layer. In this paper, our proposed architecture is L1(30×4)-L3(60×3)-L4(60×3)-L6(90×2)-L7(90×2)-L9(80×1). Each value indicates the number of neurons used in the corresponding layer. For example, the number of L1(30×4) means that there are 4 subnets in L1 and each subnet has 30 neurons. The distortions of input data and weight are 0.1 and 0.05, respectively. The momentum is set as 0.9 and we update all the weights after learning each sequence. We adopt the tanh function as the activation function of PHRNN. In the training stage, the physical correlations and constraints among facial parts, as well as weight noises are helpful to restrain the overfitting problem.

B. MSCNN for Global Static Features

As shown at the bottom part of Figure 1, 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 [11]. Thus, we adopt the ReLU function as the activation function of neurons, which has achieved better performance than the sigmoid function.

Researchers design various encouraging models to extract powerful features in recent years [12], [14]. However, a small database is an large impediment to successfully recognize the facial expression by applying these models directly. Meanwhile, most of these models use the recognition signal as supervision, which is useful to 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 variations of identical expressions. In order to apply deep models to extract powerful features, we employ an additional expression verification signal which is not only helpful to enlarge between-expression differences, but also can reduce the within-expression variations. The two signals correspond to two different loss functions which work together to push our model to focus on expression information, rather than identities, illuminations, ages and son on. Figure 2 shows the process for training the MSCNN. We will prove the advantage of two signals and visualize our learned features in Sections IV-C.

Multi-signal with different loss functions. Namely, our proposed MSCNN learns features under the supervision of two signals. The first is expression recognition signal, which can classify each face image into different expressions. For the objective function, we train the network by minimizing the cross-entropy loss, which is defined as:

$$ReLoss(p, q) = -\sum_{x} p(x) log\, q(x), \tag{10}$$

where x is the spatial feature vector, $p(x)$ is the true distribution, and $q(x)$ is the predicted probability of facial expressions. The MSCNN can learn features with large between-expression variations based on the recognition signal.

The other is expression verification signal, which is effective to reduce the variations of within-expression features. We adopt the loss function based on the L2 norm [40], which is denoted as:

$$VeLoss(x_i, x_j, \theta_{ij}) = \tag{11}$$
$$\begin{cases} \frac{1}{2} \left\| f(x_i) - f(x_j) \right\|_2^2 & if\, \theta_{ij} = 1 \\ \frac{1}{2} \max\left(0,\, \delta - \left\| f(x_i) - f(x_j) \right\|_2\right)^2 & if\, \theta_{ij} = 0 \end{cases}$$

where x_i and x_j are two input facial images, and f_i and f_j are their features extracted from the fully-connected layer. $\theta_{ij} = 1$ means that x_i and y_j are from the same facial expression, and we enforce the features f_i and f_j to be close. $\theta_{ij} = 0$ means the two input images have different expressions. In this case, we push their features apart. δ is the size of the margin. This formula requires the distance of different-expression features is larger than δ. In the training phase, a hyperparameter k is utilized to balance the recognition and

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2689999, IEEE Transactions on Image Processing

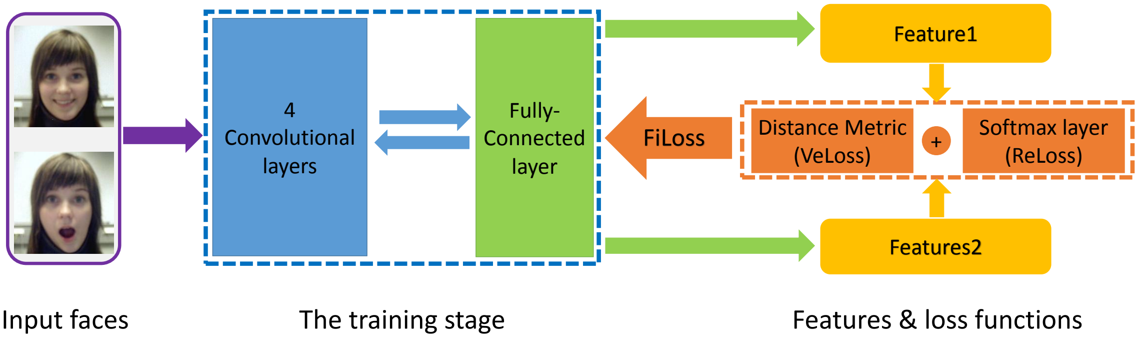


Fig. 2. Our proposed MSCNN for static facial information. In the training phase, a pair of facial images are sent into our model, and expression features are produced from the neuron activation of the fully-connected layer. The recognition and verification signals corresponds to two different loss functions, which can be combined to update all weights of our model. In the testing phase, our proposed MSCNN takes one current frame as input as shown in the bottom of Figure 1.

verification signals. We will study the hyperparameter k in the following section.

The output of the softmax layer is the probability distribution of n different facial expressions.

$$p_i = \frac{e^{y_i}}{\sum_{j=1}^{n} e^{y_j}} \tag{12}$$

$$y_j = \sum_{i=1}^{m} x_i \cdot w_{i,j} + b_j \tag{13}$$

where m denotes the number of neuron in the fully connected layer. x_i, b_j and $w_{i,j}$ are the input map, the bias and weights between two layers, respectively.

Implementation details and overfitting problem. Faces are detected with a facial detection algorithm [41]. To eliminate the impact of interference information, we crop and align the facial images according to the face location. Then we resize the facial images to the size of 64×64×1. In order to expand the training samples, we randomly crop a 60×60 patch from one image that is put into our CNN. Moreover, the patches are randomly flipped. In this way, there are 32 possible samples in one image.

The use of two signals expands the total amount of information from N images to about $N \times N$ pairs of samples. More specifically, every pairs of samples are put into our model as a whole in the training stage. This is beneficial to alleviate the overfitting problem. Our model is trained by back-propagation with the recognition loss function and verification loss function. Weights are initialized by a Gaussian distribution with zero mean and a standard deviation of 0.01. The biases are initialized as 1. In all layers, the momentum is set as 0.9 and the weight decay is set as 0.005. The dropout learning is used for the fully-connected layers which is set to be 0.5. In each iteration, we update the weights after learning one mini-batch with a size of 128.

C. Model Fusion

In our Spatial-Temporal Networks, the temporal network (PHRNN) and spatial network (MSCNN) are combined by a fusion function:

$$O(x) = \sum_{i=1}^{2} \alpha_i (A_i(x) + P_i(x)) \tag{14}$$

where $P_i(x)$ $(0 < P_i(x) < 1)$ is the predicted probability of expressions. It is the output of the softmax layers in PHRNN and MSCNN. $P_0(x)$ comes from PHRNN and its formula is Equation (8), while $P_1(x)$ comes from MSCNN and its formula is Equation (12).

A_i is the predicted sorting of expression classes. It is formulated as:

$$A_i(x_1), \ldots, A_i(x_n) = sort(P_i(x_1), \ldots, P_i(x_n)) \tag{15}$$

where n denotes the total number of expression classes. n expression classes are sorted based on the value of $P_i(x)$. In other words, a bigger $P_i(x)$ corresponds to a bigger $A_i(x)$ $(A_i(x) \in \{1, 2, ..., n\})$.

The meaning of the Equation (14) is that when our models estimate facial expressions, they give priority to the predicted sorting. If Spatial-Temporal Networks cannot decide the expression based on the value of $A_i(x)$, we will compare $P_i(x)$.

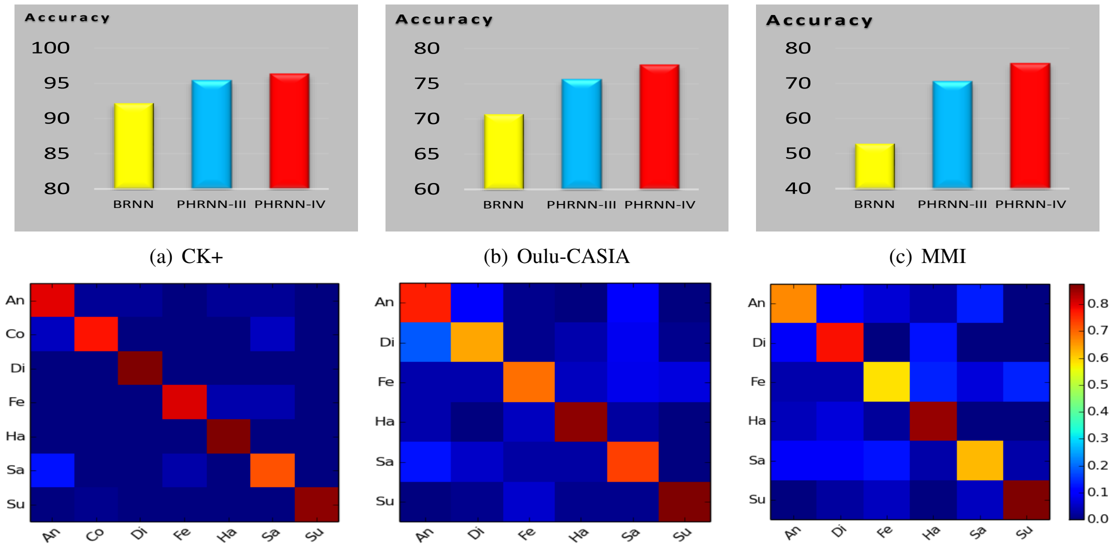
The inspiration we choose this function is that our proposed models belong to two kinds of networks with different levels of sensitivity, so the predicted sorting of the expressions in different networks is more worthy to be considered than the absolute probability. In addition, another widely used method is to average probability in different networks. We find that it has inferior performance than the proposed function in our experiments. α_i is to balance the outputs of different models. In our experiments, we test the model to find that we can achieve the best performance when α_i is set to 0.5. So we set α_i to 0.5 in all the experiments.

IV. Experiments

We assess the performance of our models on three widely used databases, namely CK+, Oulu-CASIA, and MMI. The following shows the details of experiments and results.

A. Databases and Protocols

The CK+ database. As an extended version of Cohn-Kanade (CK) [42], the CK+ database includes 123 subjects with 593 sequences. Among these sequences, 327 of them are labeled with seven emotion labels (anger, contempt, disgust, fear, happiness, sadness, and surprise). Each of the expression



(a) CK+ (b) Oulu-CASIA (c) MMI

(d) CK+ (e) Oulu-CASIA (f) MMI

Fig. 3. (a),(b)and(c) show the comparison of accuracy with different temporal networks on three databases, respectively. BRNN (yellow), PHRNN-III (blue) and PHRNN-IV (red) represent three kinds of BRNN. (d),(e)and(f) show the confusion matrices of our proposed PHRNN-IV on three databases, respectively.

sequences reflects the expression from the neutral emotion to the apex of the emotion. Although there are several validation methods on the database, we employ the most popular 10-fold validation protocol. We divide the sequences into 10 groups in ID ascending order which is the same as the previous works [6], [7], [43], [44]. In each time, nine groups are selected as training sequences and the rest one is used for testing.

The Oulu-CASIA database. The Oulu-CASIA database [25] consists of six expressions (anger, disgust, fear, happiness, sadness and surprise) from 80 subjects of 23 to 58 years old. Facial expressions are taken under three different illumination conditions: normal, weak and dark, and each illumination has 480 sequences (80 subjects with six expressions). All expression sequences begin at the neutral emotion and end with the peak of the emotion. In our experiments, we evaluate our model under the normal illumination condition [6], [7], [24]. Similar to the CK+ database, we adopt the 10-fold cross-validation strategy.

The MMI database. The MMI database [27] is composed of only 30 subjects aged from 19 to 62, of whom have both sexes and different ethnicity. In this database, 205 sequences have been labeled with six facial emotions. Different to the CK+ database and the Oulu-CASIA database, each of the sequences reflects the whole dynamic facial expression with the neutral, apex and offset phases. In this paper, we conduct our experiments on all of the 205 sequences with 10-fold cross-validation, which is the same as [6], [7], [45]. MMI is a more challenging database than the CK+ and the Oulu-CASIA. Firstly, some subjects wear accessories, such as glasses and headcloth. Secondly, the size of the database is small which is a major impediment to apply the deep learning method. Similarity to [7], we reduce the number of neurons in every layer to the half in the experiments due to the small database.

B. Comparison with Different Temporal Networks (PHRNN)

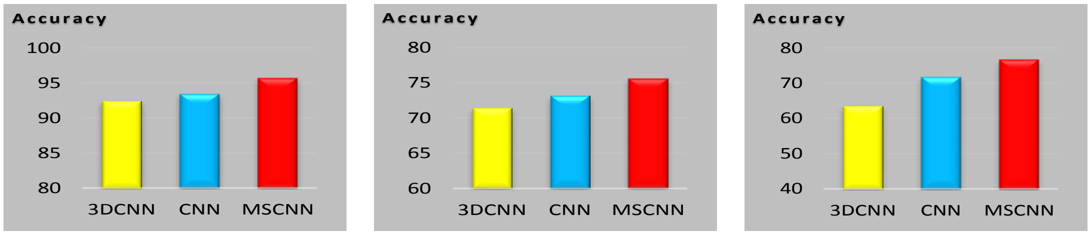
Data preprocessing. Our temporal network (PHRNN) takes facial landmarks as input. In our experiments, we detect them with the SDM algorithm [46] which can extract 49 facial landmarks including two eyebrows, two eyes, a nose and a mouth. These landmarks are detected in order. The position and region of local parts can be determined via this method, and because of this, the number of landmarks in different selected local parts are determined. Specially, the number of landmarks in eyebrows, nose, eyes and mouth are 10, 9, 12 and 18, respectively.

Facial expressions are independent of its absolute spatial position. In order to alleviate the impact of the absolute coordinate, we normalize the facial expression landmarks to a unified coordinate system. Given that the bridge of the nose is a stable point in human face, we choose its center as the origin of the new system which is formulated as:

$$\mathcal{C} = \frac{1}{4}(L_{Nose-1} + L_{Nose-2}+L_{Nose-3}+L_{Nose-4}) \qquad (16)$$

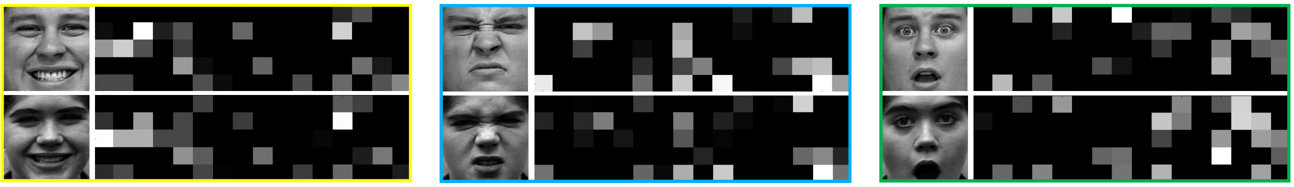
where $L_{Nose-\mathrm{i}}$ is the i-th landmark of the nose bridge. All positions of the facial landmarks are then updated by subtracting $\mathcal{C}$. Finally, the landmarks are divided into four parts and fed into PHRNN. A more challenging problem is out-of-plane rotations, which can change motion amplitudes in two-dimensional images. However, the dynamic evolution is still hidden in relative movements of facial structure and relative motion tendencies do not change. Our networks are proposed to capture facial expression information based on these relative movements. Adding several out-of-line faces to the training set is a good way to alleviate the effect caused by out-of-plane rotations.

Different temporal networks. The structure and parameters of our temporal network is L1(30×4)-L3(60×3)-L4(60×3)-L6(90×2)-L7(90×2)-L9(80×1) (PHRNN-IV), and its meaning is explained in Section III-A. We compare our model



(a) CK+ (b) Oulu-CASIA (c) MMI

Fig. 5. Comparison of accuracy with different CNN structures (3DCNN, CNN and MSCNN) on three databases.



(a) Happiness (b) Disgust (c) Surprise

Fig. 6. Examples of the learned features from the spatial network (MSCNN) on the CK+ database. From left to right are (a) happiness, (b) disgust and (c) surprise, respectively. The features show condition of neurons in the last fully-connected layer. Brighter squares indicate higher values. One image has 80 squares because the last fully-connected layer has 80 neurons.

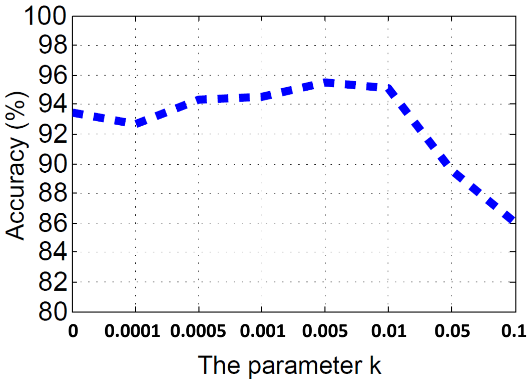


Fig. 4. Facial expression recognition accuracy of our proposed model with different parameter k.

with another two BRNN. One model is L1(120)-L3(180)-L4(180)-L6(180)-L7(180)-L9(80) (BRNN), which takes the whole facial landmarks as the input like traditional structures. The other model is L1(30×3)-L3(60×3)-L4(60×3)-L6(90×2)-L7(90×2)-L9(80×1) (PHRNN-III), which divides the facial landmarks into three parts including an eyebrow-eye, a nose and a mouth. The reason we compare our model with PHRNN-III is that eyebrows are closely related to eyes in facial variations.

Comparison. Figures 3(a), 3(b) and 3(c) show the results of different temporal networks on three databases. **PHRNN-III** (with three local parts) and **PHRNN-IV** (with four local parts) obtain a higher average accuracy than **BRNN** (with one part), which proves the effectiveness of part-based hierarchical methods for modeling dynamic variations. Dividing facial landmarks into several parts based on facial physical structure and feeding them into different subnets is beneficial to extract local and global information, which is helpful to model the dynamically evolutional properties of facial expression.

As another important element, the movement of facial contour is closely related to the motion of mouth, which can be indirectly represented by the mouth's motion. In order to reduce the calculation, we do not additionally analyze the dynamical evolution of facial contour. And in our experiments, we actually test the extra model include the facial contour, yet the performance was not improved. Thus, we chose the PHRNN-IV as our temporal networks in our following experiments. The confusion matrices of PHRNN-IV on the three expression databases are shown in Figures 3(d), 3(e) and 3(f).

C. Evaluation of the Spatial Network (MSCNN)

The deep spatial network (MSCNN) takes the detected facial images as input. In this paper, we choose the last frame (CK+), last frame (Oulu-CASIA) and middle frame (MMI) of sequences on three different databases, respectively. In the training phase, we can obtain the $ReLoss$ based on recognition signals and the $VeLoss$ based on verification signals.

Balancing. These two loss functions are weighted by a hyperparameter k to calculate the final loss $FiLoss$:

$$FiLoss(x) = \mathrm{Re}Loss(x) + k \times VeLoss(x) \qquad (17)$$

We update the weight according to the $FiLoss$ which is shown in Figure 2.

In order to balance the recognition and the verification signals, we investigate the weight by varying k from 0 to $+\infty$. When $k = 0$, only the recognition signal remains. With the increasing of k, the verification signal plays a more and more important role in our spatial network. According to Figure 4, we find that the performance of our network becomes better as k increases, and achieves the highest accuracy when $k = 0.005$. After that the model's performance begins to drop. Therefore k is set as 0.005 when we conduct our experiments with 10-fold cross-validation.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIP.2017.2689999, IEEE Transactions on Image Processing

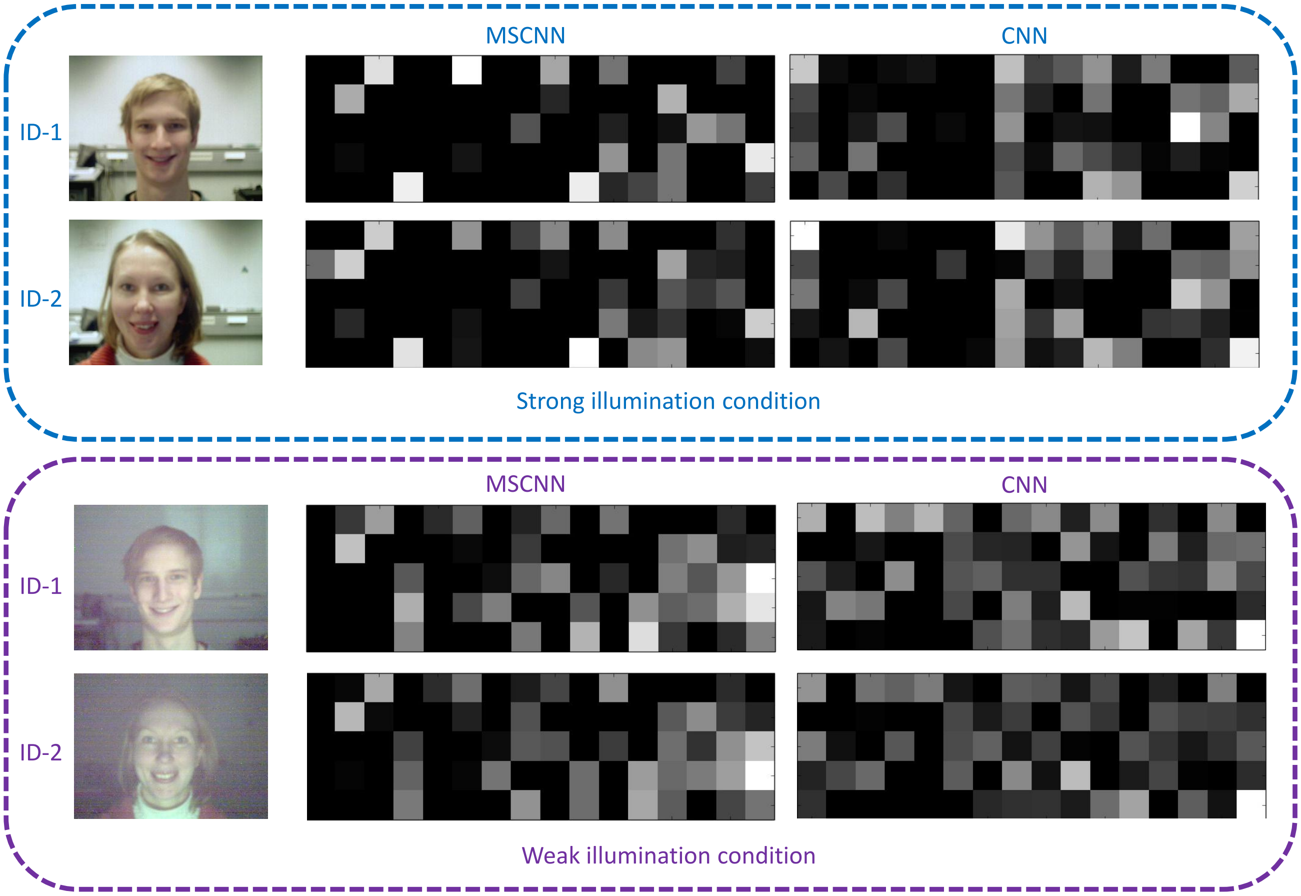


Fig. 7. Visualizations of the top level learned features for similar and different subjects under different conditions but similar facial expression for both the CNN and our proposed system. The two subjects come from two different conditions(strong illumination and wrong illumination), and both of them have a similar facial expression (happiness). The images on the top two rows are under strong illumination condition, while the images on the bottom two rows are under weak illumination condition. Models are trained under strong illumination on the Oulu-CASIA database.

Because these two loss functions are two different types, an idea is to normalize them and then calculate the final loss $FiLoss$, which is used to update our proposed model. While the two loss functions will be balanced by the parameter k after normalization, an alternative method is to tune the parameter k directly to calculate the final loss function. The parameter k is tuned based on the accuracy of facial expression recognition. Specifically, the accuracy varies as the change of parameter k. When our model achieves the highest accuracy, the parameter k at this moment is the selected one.

Comparison. Figure 5 shows the performance of the traditional **CNN** structure and our proposed **MSCNN** on three databases. We also compare with a typical **3DCNN** structure proposed by Liu *et al.* [37]. The comparable results between CNN and 3DCNN show that the still frame of facial expression has also strong discrimination. The performance of the MSCN-N is better than the CNN, and this shows that the two kinds of signals are useful to learn powerful features. Table I shows the effect of the verification signal and the sorting item. Therefore, we conduct the following experiments with the verification signal and sorting item.

Visualization. Figure 6 shows the learned features from MSCNN. We can find that the different subjects with the identical expression have very common activated neurons while the identical subject has large variations of activated neurons if he/she is in different emotions. This phenomenon demonstrates that our model can learn expression itself implied in facial images. The two signals corresponding to two kinds of loss functions are helpful to increase the variations of different expressions and reduce the difference among the identical expressions, which can be force our model to focus on expression information when same expression is presented in different identifies, genders and so on.

TABLE I
COMPARISONS OF DIFFERENT MODELS ON THE CK+ DATABASE.

Model	Explanation	Accuracy
CNN	without the verification signal	93.4%
MSCNN	with the verification signal	95.7%
PHRNN-MSCNN	without the sorting term	96.7%
PHRNN-MSCNN	with the sorting term	98.5%

Figure 7 shows visualizations of the top level learned features from both the CNN and MSCNN under strong and weak illumination condition. According to this figure, the activated neurons of identical expression in our proposed model are more similar than those in the baseline CNN.

TABLE II
COMPARISONS OF DIFFERENT METHODS ON THE CK+ DATABASE. THE PREVIOUS BEST ACCURACY IS OBTAINED BY JUNG *et al.* [7].

Method	Descriptor	Accuracy
Ptucha *et al.* [44]	MSR	91.4%
Klaser *et al.* [5]	HOG 3D	91.44%
Walecki *et al.* [43]	VSL-CRF	93.9%
Liu *et al.* [6]	STM-ExpLet	94.19%
Liu *et al.* [37]	3DCNN	85.9%
Liu *et al.* [37]	3DCNN-DAP	92.4%
Jung *et al.*[7]	CNN	91.44%
Jung *et al.*[7]	DNN	92.35%
Jung *et al.*[7]	CNN-DNN	97.25%
Spatial Network	**MSCNN**	**95.54%**
Temporal Network	**PHRNN**	**96.36%**
Spatio-Temporal Networks	**PHRNN-MSCNN**	**98.50%**

TABLE III
COMPARISONS OF DIFFERENT METHODS ON THE OULU-CASIA DATABASE. THE PREVIOUS BEST ACCURACY IS OBTAINED BY JUNG *et al.*[7].

Method	Descriptor	Accuracy
Klaser *et al.* [5]	HOG 3D	70.63%
Zhao *et al.* [25]	AdaLBP	73.54%
Liu *et al.* [6]	STM-ExpLet	74.59%
Guo *et al.* [24]	Atlases	75.52%
Jung *et al.*[7]	CNN	74.38%
Jung *et al.*[7]	DNN	74.17%
Jung *et al.*[7]	CNN-DNN	81.46%
Spatial Network	**MSCNN**	**77.67%**
Temporal Network	**PHRNN**	**78.96%**
Spatio-Temporal Networks	**PHRNN-MSCNN**	**86.25%**

TABLE IV
COMPARISONS OF DIFFERENT METHODS ON THE MMI DATABASE. THE PREVIOUS BEST ACCURACY IS OBTAINED BY LIU *et al.* [6].

Method	Descriptor	Accuracy
Klaser *et al.* [5]	HOG 3D	60.89%
Scovanner *et al.* [23]	3D SIFT	64.39%
Zhong *et al.* [45]	CSPL	73.53%
Liu *et al.* [6]	STM-ExpLet	75.12%
Liu *et al.* [37]	3DCNN	53.2%
Liu *et al.* [37]	3DCNN-DAP	63.4%
Jung *et al.*[7]	CNN	62.45%
Jung *et al.*[7]	DNN	59.02%
Jung *et al.*[7]	CNN-DNN	70.24%
Spatial Network	**MSCNN**	**77.05%**
Temporal Network	**PHRNN**	**76.17%**
Spatio-Temporal Networks	**MSCNN-PHRNN**	**81.18%**

TABLE V
CONFUSION MATRIX OF SPATIO-TEMPORAL NETWORKS ON THE CK+ DATABASE.

	An	Co	Di	Fe	Ha	Sa	Su
An	**100**	0	0	0	0	0	0
Co	0	**100**	0	0	0	0	0
Di	0	0	**100**	0	0	0	0
Fe	0	0	0	**100**	0	0	0
Ha	0	0	0	0	**100**	0	0
Sa	10.71	0	0	3.57	0	**85.71**	0
Su	0	1.23	0	0	0	0	**98.77**

TABLE VI
CONFUSION MATRIX OF SPATIO-TEMPORAL NETWORKS ON THE OULU-CASIA DATABASE.

	An	Di	Fe	Ha	Sa	Su
An	**91.25**	3.75	1.25	0	3.75	0
Di	10	**82.5**	1.25	1.25	5	0
Fe	3.75	2.5	**85**	5	1.25	2.5
Ha	3.75	0	5	**88.75**	2.5	0
Sa	15	5	0	1.25	**78.75**	0
Su	0	0	7.5	0	1.25	**91.25**

TABLE VII
CONFUSION MATRIX OF SPATIO-TEMPORAL NETWORKS ON MMI.

	An	Di	Fe	Ha	Sa	Su
An	**80**	3.33	0	3.33	13.33	0
Di	6.25	**81.25**	0	12.5	0	0
Fe	13.79	0	**55.17**	6.9	3.45	20.69
Ha	2.38	9.52	0	**88.1**	0	0
Sa	3.13	3.13	9.38	3.13	**81.25**	0
Su	5	2.5	2.5	0	0	**90**

Our models are trained under strong illumination condition, while it is able to extract the similar feature of the identical expression under weak illumination condition. This means that the focus on the facial expression recognition. While the features extracted from identical expression under different illumination conditions are less similar than those under the same illumination condition. This means that the illumination can affect the feature extraction if models are not trained under different illumination conditions. It is worth noting that all images comes from the Oulu-CASIA database. The accuracies of previous state-of-the-art methods [7], [6], [24] on this database is lower than the results on CK+. Their accuracies (Tables III) are achieved under a strong illumination condition.

D. Results and Analysis

Accuracy. Tables II, III and IV compare the performance of our models with current state-of-the-art methods on three databases. For the CK+ database, the previous best algorithm based on traditional methods for facial expression recognition is STM-ExpLet [6] which achieves the 94.19% accuracy. Recently, Jung *et al.* [7] propose a DTAGN model to overcome the previous hand-crafted method. They utilize a CNN and DNN to capture the dynamic variations and obtain 97.25%. Table II shows that our designed Spatial-Temporal Networks (PHRNN-MSCNN) achieve satisfactory performance which outperform the state-of-the-art method. For the Oulu-CASIA and MMI databases, our proposed models also significantly outperform the previous best methods. All of the comparative methods adopt the same 10-fold cross validation.

PHRNN and MSCNN. It's worth mentioning that each of PHRNN and MSCNN outperforms all the methods on three databases, excepting the DTAGN (joint CNN and DNN) proposed by Jung *et al.* [7]. This is because that our PHRNN has the ability to model the dynamically evolutional properties and MSCNN can focus on expression itself better based on different loss functions. In addition, the Spatial-Temporal Networks consider the partial-whole, geometry-appearance and dynamic-static information simultaneously. Thus, PHRNN and MSCNN can complement each other and boost the performance.

Confusion matrices. The confusion matrices for Spatial-Temporal Networks on three databases are shown in Tables V, VI and VII. We can see that our models perform well on anger, disgust, happiness and surprise, but have relatively poor performance for recognizing fear and sadness. The reasons come from two aspects: the motion of facial critical areas for fear and sadness is slight which creates more difficulties for PHRNN to distinguish the small dynamic variations from videos, and the appearances of facial images are similar to other emotions such as anger which is an impediment for the MSCNN to distinguish static information from still frames.

Computational efficiency. Finally, we evaluate the speed of our proposed models on the CK+ database. The platform is a normal desktop PC with Intel(R) Core(TM) i7-4770K CPU @ 3.50GHz and GeForce GTX TITAN GPU. The spatial network (MSCNN) runs at 909 FPS (Frame per Second) on a GPU core and 333 FPS using a single CPU. The temporal network (PHRNN) runs at 367 sequences per second on a single CPU core and each sequence has about 16 frames. Thus, it is possible to apply the proposed method for real-time frame-by-frame applications, even by combining them together in practical applications

V. Conclusion and Future Work

In this paper, we have proposed Evolutional Spatial-Temporal Networks to extract multiple kinds of features for facial expression recognition. 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.

As we have analyzed in the confusion matrices on the CK+, Oulu-CASIA and MMI databases, it is difficult to capture spatial-temporal information of expressions with slight motion. In the future, we will consider to develop more powerful structures to model the motion of facial critical areas and utilize specific methods, *e.g.*, metric learning, to analyze these emotions.

Acknowledgment

This work is jointly supported by National Key Research and Development Program of China (2016YFB1001000), National Natural Science Foundation of China (61525306, 61633021, 61420106015). This work is also supported by Strategic Priority Research Program of the CAS (Grant XD-B02070001), Youth Innovation Promotion Association CAS (2006121) and Beijing Nova Program.

References

[1] A. Vinciarelli, M. Pantic, and H. Bourlard, "Social signal processing: Survey of an emerging domain," *Image and Vision Computing*, vol. 27, no. 12, pp. 1743–1759, 2009.

[2] P. Lucey, J. Cohn, S. Lucey, I. Matthews, S. Sridharan, and K. M. Prkachin, "Automatically detecting pain using facial actions," in *International Conference on Affective Computing and Intelligent Interaction (ACII)*, IEEE, 2009.

[3] M. Pantic and L. J. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 22, no. 12, pp. 1424–1445, 2000.

[4] C. Shan, S. Gong, and P. W. McOwan, "Facial expression recognition based on local binary patterns: A comprehensive study," *Image and Vision Computing*, vol. 27, no. 6, pp. 803–816, 2009.

[5] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3d-gradients," in *British Machine Vision Conference (BMVC)*, 2008.

[6] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014.

[7] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognitin," in *The IEEE International Conference on Computer Vision (ICCV)*, IEEE, 2015.

[8] A. Graves *et al.*, *Supervised sequence labelling with recurrent neural networks*, vol. 385. Springer, 2012.

[9] A. Graves, A.-r. Mohamed, and G. Hinton, "Speech recognition with deep recurrent neural networks," in *The IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, 2013.

[10] J. Donahue, L. Anne Hendricks, S. Guadarrama, M. Rohrbach, S. Venugopalan, K. Saenko, and T. Darrell, "Long-term recurrent convolutional networks for visual recognition and description," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[12] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[13] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014.

[14] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *International Conference on Learning Representations (ICLR)*, 2015.

[15] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2016.

[16] Z. Zeng, M. Pantic, G. Roisman, T. S. Huang, *et al.*, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 1, pp. 39–58, 2009.

[17] C. Shan, S. Gong, and P. W. McOwan, "Conditional mutual infomation based boosting for facial expression recognition.," in *British Machine Vision Conference (BMVC)*, 2005.

[18] M. S. Bartlett, G. Littlewort, I. Fasel, and J. R. Movellan, "Real time face detection and facial expression recognition: Development and applications to human computer interaction.," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2003.

[19] M. Pantic and L. J. Rothkrantz, "Facial action recognition for facial expression analysis from static face images," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 34, no. 3, pp. 1449–1461, 2004.

[20] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 29, no. 6, pp. 915–928, 2007.

[21] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas, "Facial expression recognition using encoded dynamic features," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2008.

[22] S. W. Chew, P. Lucey, S. Lucey, J. Saragih, J. F. Cohn, and S. Sridharan, "Person-independent facial expression detection using constrained local models," in *The IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, IEEE, 2011.

[23] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th International Conference on Multimedia*, ACM, 2007.

[24] Y. Guo, G. Zhao, and M. Pietikäinen, "Dynamic facial expression recognition using longitudinal facial expression atlases," in *European Conference on Computer Vision (ECCV)*, Springer, 2012.

[25] M. Taini, G. Zhao, S. Z. Li, and M. Pietikäinen, "Facial expression recognition from near-infrared video sequences," in *International Conference on Pattern Recognition (ICPR)*, 2008.

[26] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *The IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, 2010.

[27] M. Valstar and M. Pantic, "Induced disgust, happiness and surprise: an addition to the mmi facial expression database," in *Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect*, 2010.

[28] A. Graves, M. Liwicki, S. Fernández, R. Bertolami, H. Bunke, and J. Schmidhuber, "A novel connectionist system for unconstrained handwriting recognition," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 31, no. 5, pp. 855–868, 2009.

[29] A. Graves and J. Schmidhuber, "Offline handwriting recognition with multidimensional recurrent neural networks," in *Advances in Neural Information Processing Systems (NIPS)*, 2009.

[30] G. Lefebvre, S. Berlemont, F. Mamalet, and C. Garcia, "Blstm-rnn based 3d gesture classification," in *Artificial Neural Networks and Machine Learning–ICANN*, Springer, 2013.

[31] A. Graves, C. Mayer, M. Wimmer, J. Schmidhuber, and B. Radig, "Facial expression recognition with recurrent neural networks," in *Proceedings of the International Workshop on Cognition for Technical Systems, Munich, germany*, 2008.

[32] H. Vadapalli, H. Nyongesa, and C. Omlin, "Recurrent neural networks for facial action unit recognition from image sequences," in *Proceedings of the 2010 International Conference on Image Processing, Computer Vision, & Pattern Recognition*, 2010.

[33] S. E. Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *ACM International Conference on Multimodal Interaction*, 2015.

[34] Y. Sun, X. Wang, and X. Tang, "Deep learning face representation from predicting 10,000 classes," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2014.

[35] Y. Sun, Y. Chen, X. Wang, and X. Tang, "Deep learning face representation by joint identification-verification," in *Advances in Neural Information Processing Systems (NIPS)*, 2014.

[36] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[37] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Asian Conference on Computer Vision (ACCV)*, Springer, 2014.

[38] A. Graves and N. Jaitly, "Towards end-to-end speech recognition with recurrent neural networks," in *International Conference on Machine Learning (ICML)*, 2014.

[39] M. Schuster and K. K. Paliwal, "Bidirectional recurrent neural networks," *Signal Processing, IEEE Transactions on*, vol. 45, no. 11, pp. 2673–2681, 1997.

[40] H. Mobahi, R. Collobert, and J. Weston, "Deep learning from temporal coherence in video," in *Proceedings of the 26th Annual International Conference on Machine Learning*, ACM, 2009.

[41] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2013.

[42] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *The IEEE International Conference on Automatic Face and Gesture Recognition and Gesture Recognition (FG)*, IEEE, 2000.

[43] R. Walecki, O. Rudovic, V. Pavlovic, and M. Pantic, "Variable-state latent conditional random fields for facial expression recognition and action unit detection," *Proceedings of IEEE*, pp. 1–8.

[44] R. Ptucha, G. Tsagkatakis, and A. Savakis, "Manifold based sparse representation for robust expression recognition without neutral subtraction," in *The IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, IEEE, 2011.

[45] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2012.

[46] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, IEEE, 2013.

Kaihao Zhang received the M. Eng. degrees in Computer Application Technology from the University of Electronic Science and Technology of China, Chengdu, China, in 2016. Before he went to Australia, he had been doing research at the Center for Research on Intelligent Perception and Computing, National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. Currently, he is working toward the PhD degree at the College of Engineering and Computer Science, the Australian National University, Canberra, ACT, Australia. His research interests focus on face detection, face verification and facial expression recognition with deep learning.

Yongzhen Huang received his B.E. degree from Huazhong University of Science and Technology (HUST) in 2006, and his Ph.D. degree from Institute of Automation, Chinese Academy of Sciences (CASIA) in 2011. Then he joined National Laboratory of Pattern Recognition (NLPR) as an Assistant Professor in July 2011, and became an Associated Professor since Nov. 2013. His research interests include computer vision, pattern recognition and machine learning. He has published one book and more than 60 papers in international journals and conferences such as IEEE TPAMI, IJCV, IEEE TMSC-B, IEEE TCSVT, TMM, CVPR, ICCV, NIPS.

Yong Du received the BS degree in electronic information science and technology, the MS degree in signal and information processing from Chengdu University of Technology, China, in 2010 and 2013, respectively. He is currently working toward the PhD degree in pattern recognition and intelligent system at the Center for Research on Intelligent Perception and Computing, National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His current research interests mainly include video analysis and action recognition with deep learning.

Liang Wang (SM'09) received the PhD degree in Pattern Recognition and Intelligent System from the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CAS), China, in 2004. After graduation, he has worked as a Research Assistant at the Imperial College London, United Kingdom and Monash University, Australia, and a Research Fellow at the University of Melbourne, Australia, respectively. Before he returned back to China, he was a Lecturer with the Department of Computer Science, University of Bath, United Kingdom. Currently, he is a Professor of Hundred Talents Program of CAS at the Institute of Automation, Chinese Academy of Sciences, P. R. China. His major research interests include machine learning, pattern recognition, computer vision, multimedia processing, and data mining. He has widely published at highly-ranked international journals such as IEEE TPAMI, IEEE TIP, IEEE TKDE, IEEE TCSVT, IEEE TSMC, CVIU, and PR, and leading international conferences such as CVPR, ICCV and ICDM. He has obtained several honors and awards such as the Special Prize of the Presidential Scholarship of CAS and the Research Commendation from University of Melbourne in recognition of Excellent Research. He is currently a Senior Member of IEEE, as well as a member of IEEE Computer Society, IEEE Communications Society and BMVA (British Machine Vision Association). He is serving with more than 20 major international journals and more than 40 major international conferences and workshops. He is an associate editor of IEEE Transactions on Systems, Man and Cybernetics Part B, International Journal of Image and Graphics (WorldSci), International Journal of Signal Processing (Elsevier), Neurocomputing (Elsevier), and International Journal of Cognitive Biometrics. He is a leading guest editor of 3 special issues appearing in PRL (Pattern Recognition Letters), IJPRAI (International Journal of Pattern Recognition and Artificial Intelligence) and IEEE TSMC-B, as well as a co-editor of 5 edited books. He has also co-chaired 8 international workshops.