IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 16, 2019, accepted February 18, 2019, date of publication March 5, 2019, date of current version March 26, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2901521

Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning

SHIQING ZHANG[1], XIANZHANG PAN[1], YUELI CUI[1], XIAOMING ZHAO[1], AND LIMEI LIU[2]

[1]Institute of Intelligent Information Processing, Taizhou University, Taizhou 318000, China
[2]Institute of Big Data and Internet Innovation, Hunan University of Commerce, Changsha 410205, China

Corresponding author: Shiqing Zhang (tzczsq@163.com)

This work was supported in part by the Zhejiang Provincial National Science Foundation of China under Grant LY16F020011, in part by the Major Project for National Natural Science Foundation of China under Grant 71790615, in part by the Taizhou Science and Technology Project under Grant 1802gy06, and in part by the Training Project of Taizhou University under Grant 2017PY026 and Grant 2018JQ003.

ABSTRACT One key challenging issues of facial expression recognition (FER) in video sequences is to extract discriminative spatiotemporal video features from facial expression images in video sequences. In this paper, we propose a new method of FER in video sequences via a hybrid deep learning model. The proposed method first employs two individual deep convolutional neural networks (CNNs), including a spatial CNN processing static facial images and a temporal CN network processing optical flow images, to separately learn high-level spatial and temporal features on the divided video segments. These two CNNs are fine-tuned on target video facial expression datasets from a pre-trained CNN model. Then, the obtained segment-level spatial and temporal features are integrated into a deep fusion network built with a deep belief network (DBN) model. This deep fusion network is used to jointly learn discriminative spatiotemporal features. Finally, an average pooling is performed on the learned DBN segment-level features in a video sequence, to produce a fixed-length global video feature representation. Based on the global video feature representations, a linear support vector machine (SVM) is employed for facial expression classification tasks. The extensive experiments on three public video-based facial expression datasets, i.e., BAUM-1s, RML, and MMI, show the effectiveness of our proposed method, outperforming the state-of-the-arts.

INDEX TERMS Facial expression recognition, spatio-temporal features, hybrid deep learning, deep convolutional neural networks, deep belief network.

I. INTRODUCTION

Facial expression is one of the most natural nonverbal ways for expressing human emotions and intentions. In recent years, automatic facial expression recognition (FER), which aims to analyze and understand human facial behavior, has become an increasingly active research topic in the domains of computer vision, artificial intelligence, pattern recognition, *etc*. This is because FER has many potential applications such as human emotion perception, social robotics, human-computer interaction and healthcare [1]–[5].

FER methods can be divided into two categories: video sequence-based methods (dynamic) and image-based methods (static). Most previous FER studies focus on identifying facial expressions from static facial images [1]–[4]. Although these image-based methods can effectively derive spatial information from still images, they cannot capture the temporal variability in consecutive frames in video sequences. As a dynamic event, classifying facial expression from consecutive frames in a video is more natural, since video sequences provides much more information for FER than static facial images. One key issue for video sequence-based FER methods is how to effectively encode input video sequences into an appropriate feature representation. Currently, the mainstream methods employ hand-designed feature representations, such as Gabor motion energy [6], Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [7] or Local Phase Quantization from TOP (LPQ-TOP) [8]. However, these hand-designed feature representations are low-level to discriminate dynamic facial expressions. Recently, the deep

The associate editor coordinating the review of this manuscript and approving it for publication was Tariq Ahamed Ahanger.

neural network driven feature learning representations from data may achieve better performance without requiring domain expertise [9]–[15].

Inspired by the strong feature learning ability of deep neural networks, this paper proposes a new deep neural network-based FER method in video sequences by using a hybrid deep learning model. Our hybrid deep learning model contains three deep models. The first two deep models are deep Convolutional Neural Networks (CNNs) [16], including a spatial CNN network processing static facial images and a temporal CNN network processing optical flow images. These two CNNs are separately used to learn high-level spatial features and temporal features on the divided video segments. The third deep model is a deep fusion network built with a Deep Belief Network (DBN) [17] model, which is trained to jointly learn a discriminative spatio-temporal segment-level feature representation. When finishing the joint training of a DBN, an average-pooling is applied on all the divided video segments to produce a fixed-length global video feature representation. Then, a linear Support Vector Machine (SVM) is adopted to perform facial expression classification tasks in video sequences.

It is noted that two-stream CNNs have been successfully used for video action recognition [18]. Nevertheless, in [18], a score-level scheme, which belongs to a shallow fusion method, is used to merge different features produced by two-stream CNNs. This shallow fusion method is not able to effectively model the complicated non-linear joint distribution of multiple input modalities [19]. To tackle this issue, it is desired to design deep fusion methods which leverage a deep fusion model to implement multiple meaningful feature fusion operations. Since a DBN model consists of multiple RBMs, each of which can be used to jointly learn feature representations of multiple input modalities, it may be feasible to use a DBN model as a deep fusion method to integrate different features produced by two-stream CNNs. This motivates us to develop a hybrid deep leaning method to learn video features for facial expression recognition in video sequences. Experiment results on three public video-based facial expression databases, including the BAUM-1s database [20], the RML database [21], and the MMI database [22], are presented to demonstrate the effectiveness of the proposed method on FER tasks in video sequences.

The distinct features of this paper can be summarized in two-fold: (1) We propose a hybrid deep learning model, comprising a spatial CNN network, a temporal CNN network and a deep fusion network built with a DBN model, to apply for FER in video sequences. To the best of our knowledge, it is the first time to employ a hybrid deep learning model to learn video features for FER in video sequences. (2) To deeply fuse the spatial CNN features and temporal CNN features, we employ a deep DBN model as a deep fusion network to learn a joint discriminative spatio-temporal segment-level feature representation for FER. Extensive experiments are conducted on three public video-based facial expression datasets, and experiment results demonstrate that our method outperforms the-state-of-the-arts.

The structure of this paper is organized as follows. Section 2 reviews the related work in brief. Section 3 describes our proposed method in detail. Experiment results and analysis are given in Section 4. Section 5 presents the conclusions and future work.

II. RELATED WORK

In this section, we review the recent works related to feature extraction in FER in video sequences, which uses hand-designed features and deep learning-based features.

A. HAND-DESIGNED FEATURE-BASED METHOD

For facial feature representation in static images, a variety of local image descriptors, including Local Binary Pattern (LBP) [23], Histogram of Oriented Gradient (HOG) [24], and Scale Invariant Feature Transform (SIFT) [25] have been widely used for FER. For dynamic expression recognition, these typical local features have been extended and applied to video sequences, such as LBP-TOP [7], LPQ-TOP [8], 3D-HOG [26], 3D-SIFT [27], respectively. Hayat *et al.* [28] compare the performance of various dynamic descriptors including HOG, 3D-HOG, 3D-SIFT and LBP-TOP by using bag of features framework for video-based FER, and find that LBP-TOP performs best among these dynamic descriptors. Additionally, spatio-temporal Gabor motion energy filters [6] is presented for low-level integration of spatio-temporal information on FER tasks.

Recently, some efforts have been conducted to develop more powerful spatio-temporal feature extraction methods for FER. For instance, Liu *et al.* [29] present an expressionlet-based spatio-temporal manifold descriptor which shows the superiority over traditional methods on FER tasks. Fan and Tjahjadi [30] provide a spatio-temporal feature based on local Zernike moment and motion history image for dynamic FER. Yan [31] proposes a collaborative discriminative multi-metric learning for FER in video sequences. In particular, for each video sequence they firstly calculate multiple feature descriptors such as 3D-HOG, and geometric warp features. Then, these extracted multiple features are employed to learn multiple distance metrics collaboratively to obtain complementary and discriminative information for dynamic FER.

B. DEEP LEARNING-BASED METHOD

In recent years, deep CNNs [16], [32]–[34], composed of multiple convolution layers and pooling layers, have dominated various computer vision tasks such as image classification, object detection and face recognition. These deep CNNs extends the traditional CNN model [35] into a deep multi-layered architecture which consists of five convolution layers followed by three max-pooling layers.

One of the major drawbacks of conventional CNNs is that they are able to extract spatial relationships of input images, but cannot model the temporal relationships of them in

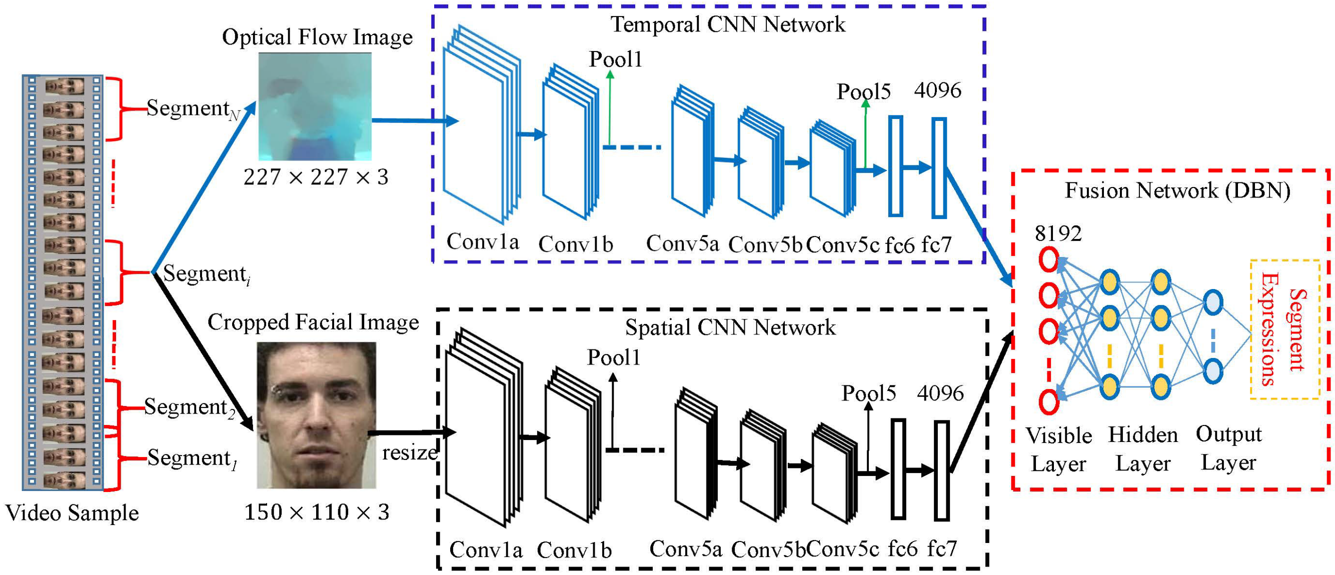


FIGURE 1. The framework of our proposed hybrid deep learning network for facial expression recognition in video sequences.

a video sequence. To solve this problem, the recently-developed 3D-CNNs [36] may present a possible solution. 3D-CNNs can extract spatio-temporal features in a video sequence by means of sliding over the temporal dimension of input data as well as the spatial dimension simultaneously. In recent years, 3D-CNNs have been used to learn spatio-temporal expression representations from successive frames in video sequences [12], [15]. In addition, a variant of 3D-CNNs is 3DCNN-DAP [14] used for dynamic FER. In 3DCNN-DAP, a constraint of Deformable Action Parts (DAP) is incorporated into the basic 3D-CNN framework. Similar to 3DCNN-DAP, Jung *et al.* [10] propose a small temporal CNN to extract temporal geometric features from facial landmark points. Although these 3D-CNNs based methods have achieved good performance on FER tasks in video sequences, but they still has a drawback. That is, these methods cannot take the deep fusion of spatio-temporal features into account simultaneously in the procedure of extracting them.

To tackle this problem, two-stream CNNs used for video action recognition [18], may present a cue. However, the used shallow fusion method in [18] based on a score-level scheme, cannot able to effectively model the complicated non-linear joint distribution of multiple input modalities. To make full use of the advantages of two-stream CNNs, we design a deep fusion network built with a deep DBN model to jointly learn the outputs of two-stream CNNs. This is our proposed hybrid deep learning model. Then, we apply this hybrid deep learning model for FER in video sequences. Experiment results on three video-based facial expression databases demonstrate the advantages of our proposed method.

III. OUR METHOD

Figure 1 shows the framework of our proposed hybrid deep learning model. As depicted in Fig.1, our method is composed of two individual channels of input streams, *i.e.*, a spatial CNN network processing static frame-level cropped facial images and a temporal CNN network processing optical flow images produced between consecutive frames. To integrate the learned spatio-temporal features represented by the outputs of fully connected layers of these two CNNs, a fusion network built with a deep DBN model is designed. In detail, our method contains four key steps: (1) generation of CNN inputs (2) spatio-temporal feature learning with CNNs (3) spatio-temporal fusion with DBNs (4) video-based expression classification. In the followings, we present the details about abovementioned four steps of our method.

A. GENERATION OF CNN INPUTS

Since CNNs require a fixed size of input data, we divide each video sample with different durations into a certain number of fixed-length segments as inputs of CNNs. This not only produces appropriate inputs of CNNs, but also augments the amount of training data to some extent.

Following in [18], the divided segment length L is set to be $L = 16$ for its good performance when using the temporal CNN network. As a result, in the latter experiments, we divide each video sample into a fixed-length segment with $L = 16$. To this end, when $L > 16$ we eliminate the first and last $(L - 16)/2$ frames. Oppositely, when $L < 16$, we simply duplicate the first and last $(16-L)/2$ frames. In this way, we make sure that each divided segment has a length of $L = 16$.

1) INPUTS OF TEMPORAL CNNs

To produce suitable inputs of temporal CNNs, we extract optical flow images between consecutive frames in a video sequence. Optical flow images represent the displacement changes of corresponding positions between consecutive frames. Following in [37], we firstly transform the values of the motion field d_x, d_y into the interval [0, 255] by

$$\tilde{d}_{x|y} = ad_{x|y} + b, \tag{1}$$

where $a = 16$, $b = 128$.

Then, the transformed flow maps are conserved as an optical flow image containing three channels, which corresponds to motion $\tilde{d}_x$, $\tilde{d}_y$ and the optical flow magnitude. In this way, we finally produce an optical flow image with size of 227 × 227 × 3. It is noted that a video segment $L = 16$ can generate 15 optical flow images as inputs of temporal CNNs, since two consecutive frames yield one optical flow image.

2) INPUTS OF SPATIAL CNNs

For inputs of spatial CNNs, we employ a cropped facial image of 150 × 110 × 3 for each frame in a video segment, as in [23]. In detail, a robust real-time face detector [38] is firstly leveraged to perform face detection to crop a facial image from each frame in a video segment. Then, in terms of the normalized distance between two eyes, a cropped image of 150 × 110 × 3 containing facial key parts, such as head, nose, mouth, *etc.*, is obtained from a facial image. Finally, we resize the cropped facial image into 227 × 227 × 3 as inputs of spatial CNNs. Note that we discard the first frame in a video segment $L = 16$, and employ the remaining 15 frames as inputs of spatial CNNs. In this case, we can make sure that the input frames of spatial CNNs in a video segment equals to that of temporal CNNs.

B. SPATIO-TEMPORAL FEATURE LEARNING WITH CNNs

As described in Fig.1, the used spatial and temporal CNNs have the same structure as the original VGG16 [16], which consists of five convolution layers (Conv1a-Conv1b, Conv2a-Conv2b, Conv3a-Conv3b-Conv3c-, $\cdots$, Conv5a, Conv5b-Conv5c), five max-pooling layers (Pool1, Pool2, $\cdots$, Pool5), and three fully connected (FC) layers (fc6, fc7, fc8). Note that fc6 and fc7 have 4096 units, while fc8 represents a class label vector which equals to data categories. Note that fc8 in VGG16 corresponds to 1000 image categories.

To realize the task of spatio-temporal feature learning with CNNs, we fine-tune the pre-trained VGG16 [16] on target video-based facial expression data. In particular, we firstly copy the existing VGG16 parameters pre-trained on large-scale ImageNet data to initialize the temporal CNN network and the spatial CNN network, respectively. Then, we replace the fc8 layer in VGG16 with a new class label vector corresponding to six facial expression categories used in our experiments. Ultimately, we individually retrain these two CNN streams by using the standard back propagation strategy. Specially, we use the back propagation technique to solve the following minimizing problem so as to update the CNN network parameters:

$$\min_{W,\theta} \sum_{i=1}^{N} H(\text{softmax}\,(W \cdot \Upsilon(a_i; \vartheta)), y_i), \tag{2}$$

where W denotes the weights of the softmax layer for the network parameters ϑ belonging to spatial CNNs or temporal CNNs. $\Upsilon(v_i; \vartheta)$ is the 4096-D output of fc7 for input data a_i. And y_i is the class label vector of the i-th segment, H is the softmax log-loss function defined as

$$H(\vartheta, y) = -\sum_{j=1}^{C} y_j \log(y_j), \tag{3}$$

where C is the total number of facial expression categories. Once both spatial CNNs and temporal CNNs are trained, the 4096-D outputs of their fc7 layers represent the learned high-level feature representations in video segments.

C. SPATIO-TEMPORAL FUSION WITH DBNs

When finishing the training of spatial CNNs and temporal CNNs, the 4096-D outputs of their fc7 layers were directly concatenated into a total 8192-D vector as inputs of the fusion network built with a deep DBN model [17], as illustrated in Fig.1. This deep DBN model is used to capture highly non-linear relationships across spatial and temporal modalities, and produce a joint discriminative feature representation for FER.

A DBN model is a multi-layered neural network structure formed by stacking a series of Restricted Boltzmann Machines (RBMs) [39], each of which is a bipartite graph. In Fig.1, two RBMs constituted by one visible layer and two hidden layers, are presented as an illustration of a DBN's structure. Here, the output layer denotes the softmax layer for classification. One key characteristic of a DBN is that it can employ multiple RBMs to learn a multi-layer generative model of input data. As a result, DBNs can effectively discover the distribution properties of input data, and learn the hierarchical feature representations of input data.

As done in [40], we use a two-step strategy to train the DBN fusion network, as described below.

(1) An unsupervised pre-training is conducted in the bottom-up way by means of a greedy layer-wise training algorithm. According to the logarithm of the probability of derivative, the weights of each RBM model is updated by

$$\Delta w = \varepsilon(< v_i h_j >_{\text{data}} - < v_i h_j >_{\text{model}}), \tag{4}$$

where ε denotes the learning rate, $< \cdot >$ represents the data expectation. v_i and h_j are the status of visual nodes and hidden nodes, respectively.

(2) A supervised fine-tuning is performed to update the network parameters with back propagation. Specially, supervised fine-tuning is realized by using the following loss function between input data and the reconstructed data.

$$L(x, x') = \left\| x - x' \right\|_2^2, \tag{5}$$

where x and x' separately denotes input data and the reconstruction data, $\| \|_2^2$ is the L2-norm reconstruction error.

D. VIDEO-BASED EXPRESSION CLASSIFICATION

After implementing the training of the DBN fusion network, the output of its last hidden layer represents the jointly learned discriminative spatio-temporal feature representations in video segments. Based on this learned segment-level features of DBNs, we then apply an average-pooling approach on all divided segments in a video sample to produce a fixed-length global video feature representation for FER. Finally, a linear SVM classifier is adopted to perform the final FER tasks in video sequences.

IV. EXPERIMENTS

To verify the performance of our proposed method on FER tasks in video sequences, FER experiments are performed on three public video-based facial expression datasets, *i.e.*, the BAUM-1s database [20], the RML database [21] and the MMI database [22].

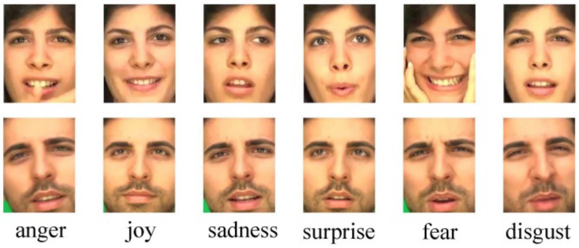


FIGURE 2. Some examples of cropped facial expression images from the BAUM-1s dataset.

A. DATASETS

1) BAUM-1s

The original BAUM-1 is a newly-developed spontaneous audio-visual face database of affective and mental states [20]. The BAUM-1 database contains not only the six basic facial expressions (joy, anger, sadness, disgust, fear, surprise) as well as boredom and contempt, but also four mental states (unsure, thinking, concentrating, bothered). It comprises of 1222 video samples collected from 31 Turkish persons. Each video frame is 720 × 576 × 3. Following in [20], we aim to identify the six basic facial expressions, which forms a small subset called the BAUM-1s dataset with 521 video samples in total. Fig.2 gives some examples of cropped facial expression images from the BAUM-1s dataset.

2) RML

The RML database [21] consists of 720 video samples collected from 8 persons. Each video frame is 720 × 480 × 3. This database has the six basic facial expressions (angry, disgust, fear, joy, sadness and surprise). Fig.3 shows some samples of cropped facial expression images from the RML database.

3) MMI

The MMI database [22] consists of 2894 video samples, out of which 213 sequences have been labeled with six basic expressions from 30 subjects aging from 19 to 62. Fig.4 provides some samples of cropped facial expression images from the MMI database.

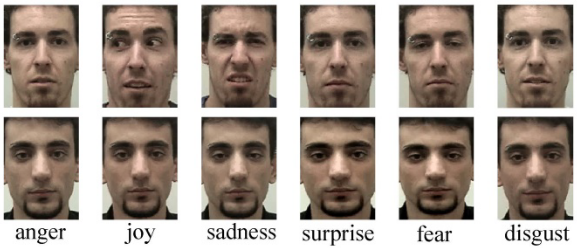


FIGURE 3. Some examples of cropped facial expression images from the RML dataset.

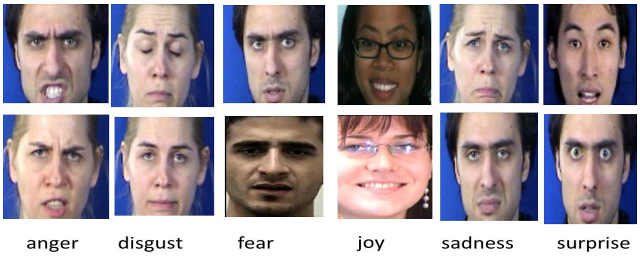


FIGURE 4. Some examples of cropped facial expression images from the MML dataset.

B. EXPERIMENT SETTINGS

When training deep neural networks, we adopt a mini-batch size of 30. The maximum number of epochs is 300 for CNNs, and 100 for DBNs, respectively. The learning rate is set to 0.001. To accelerate the training of deep models, one NVIDIA GTX TITAN X GPU with 12GB memory is employed. For all experiments we adopt subject-independent cross-validation strategy widely used in real applications. In particular, on the BAUM-1s and MMI database with more than 10 subjects, Leave-One-Subject-Group-Out (LOSGO) with five subject groups is employed, whereas on the RML database with less than 10 subjects, Leave-One-Subject-Out (LOSO) is used for experiments. Finally, we report the average recognition accuracy in all test-runs to testify the performance of all compared methods.

It is noted that we train deep models on the divided video segments so that the number of training data can be augmented. In this work, on the BAUM-1s database about 7000 segments are produced from 521 video samples. Similarily, on the RML database about 12, 000 segments are produced from 720 video samples, whereas on the MMI database, about 4000 segments are given from 213 video samples.

C. RESULTS AND ANALYSIS

We firstly evaluate the effects of deep structures of DBNs in the fusion network, since the deep structures of DBNs may greatly affects the performance of fusing spatio-temporal features. To verify the different structures of DBNs, we provide the performance of three different DBNs, including DBN-1 (8192-4096-6), DBN-2 (8192-4096-2048-6),

TABLE 1. Accuracy (%) of different structures of DBNS.

DBN structure	BAUM-1s	RML	MMI
DBN-1	48.15	68.86	66.82
DBN-2	52.73	71.52	69.88
DBN-3	55.85	73.73	71.43

TABLE 2. Accuracy (%) of different learned deep features.

Features	BAUM-1s	RML	MMI
Spatial CNN	50.96	64.58	60.45
Temporal CNN	49.14	50.31	48.66
Score-level fusion	53.04	71.94	68.35
DBN fusion	55.85	73.73	71.43

and DBN-3 (8192-4096-2048-1024-6). Table 1 presents the recognition accuracy of different structures of DBNs in the fusion network. From Table 1, we can observe that DBN-3 performs best among three different structures. In particular, DBN-3 presents an accuracy of 55.85% on the BAUM-1s dataset, 73.73% on the RML dataset, and 71.43% on the MMI dataset, respectively. This demonstrates that the deeper DBN exhibits stronger feature fusion ability based on the used multiple RBMs. In the latter experiments, in the fusion network we thus adopt DBN-3 as the default structure of the used DBN for its best performance.

To verify the advantages of fusing spatio-temporal features with DBNs, Table 2 shows the performance of four methods: the single spatial CNN features, the single temporal CNN features, the score-level fusion based on spatio-temporal CNN features, and the DBN fusion based on spatio-temporal CNN features. As shown in Table 2, we can see that the spatio-temporal CNN+DBN features, which fuse spatio-temporal CNN features with DBNs, outperform the other two features. This indicates the effectiveness of fusing spatio-temporal features by using a deep DBN. This is because DBNs are able to effectively discover the distribution properties of input spatio-temporal data, and learn the hierarchical feature representations of input spatio-temporal data.

To further present the recognition performance for each facial expression, Fig.5-7 separately show the confusion matrix of recognition results achieved by the DBN fusion network on these three datasets. It can be seen from Fig.5 that on the BAUM-1s dataset only “joy” and “sadness” are classified well with an accuracy of 88.44% and 72.39%, respectively, whereas other four facial expressions are identified badly with an accuracy of less than 35%. The results in Fig.6 demonstrate that on the RML dataset “disgust”, “sadness” and “surprise” are recognized well with an accuracy of more than 84%, whereas the remaining three facial expressions are distinguished with an accuracy of less than 80%. In Fig.7, we can see that “sadness” and “surprise” are distinguished with an accuracy of 100%, whereas the others are identified with an accuracy of less than 75%.

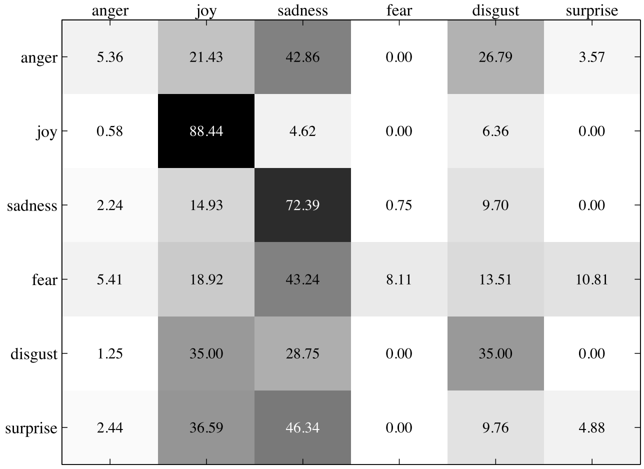


FIGURE 5. Confusion matrix of recognition results with DBNs on the BAUM-1s dataset.

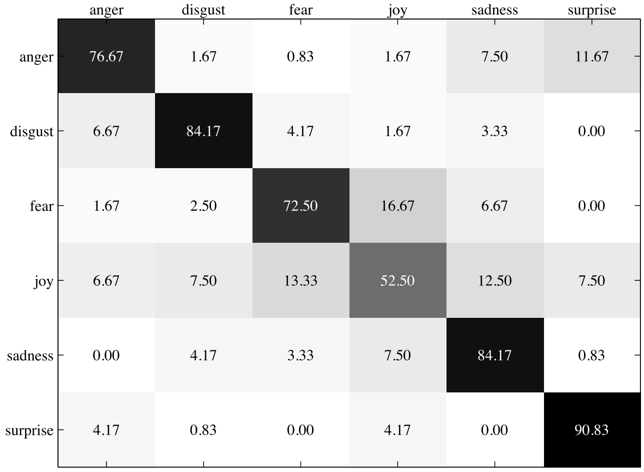


FIGURE 6. Confusion matrix of recognition results with DBNs on the RML dataset.

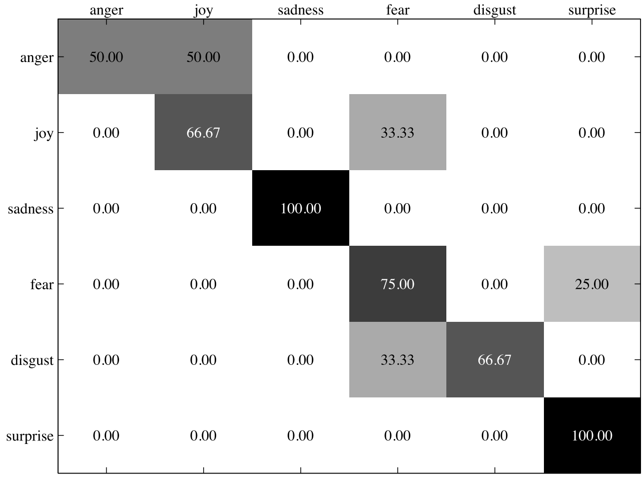


FIGURE 7. Confusion matrix of recognition results with DBNs on the MMI dataset.

Now we directly conduct a comparison with previous works on these three datasets. It is noted that these comparing works also employs subject-independent test-runs,

TABLE 3. Performance (%) comparisons of the-state-of-the-arts on the used three datasets.

Datasets	Refs.	Features	Accuracy
BAUM-1s	S Zhalehpour[20]	LPQ	45.04
	Shiqing Zhang[15]	3D-CNN	50.11
	Ours	Spatio-temporal CNN+DBN	**55.85**
RML	NED Elmadany [42]	Gabor	64.58
	Shiqing Zhang[15]	3D-CNN	68.09
	Ours	Spatio-temporal CNN+DBN	**73.73**
MMI	M. Liu [14]	3DCNN-DAP	63.40
	B. Hasani [41]	Inception-ResNet	68.51
	Ours	Spatio-temporal CNN+DBN	**71.43**

similar to ours. Table 3 provides the comparisons of the state-of-the-arts. From Table 3, it can be seen that our proposed method significantly outperforms the state-of-the-arts on these three datasets. This exhibits the superiority of our proposed method over other methods, including other deep models such as 3D-CNN [14], [15], and Inception-ResNet [41], as well as hand-designed features such as LPQ [20], and Gabor wavelets [42]. Note that 3DCNN-DAP (Deformable Action Parts) [14], 3D-CNN [15], Inception-ResNet [41] are popular spatio-temporal deep feature learning methods by using the spatial and temporal convolutions simultaneously.

V. CONCLUSION

This paper proposes a hybrid deep learning model, which consists of the spatial CNN network, the temporal CNN network, and the DBN fusion network, to apply for FER in video sequences. We implement our proposed method in two stages. (1) We employ the existing VGG16 model pre-tained on ImageNet data to individually fine-tune the spatial CNN network and the temporal CNN network on target video-based facial expression data. (2) To deeply fuse the learned spatio-temporal CNN features, we train a deep DBN model to jointly learn discriminative spatio-temporal features. Experiment results on three public video-based facial expression datasets, *i.e.*, BAUM-1s RML, and MMI, demonstrate the advantages of our proposed method.

In future, we will extend our work to practical applications. For instance, it is challenging to develop a real-time FER system based on our proposed method. In addition, it is also interesting to explore deep compression of deep models so as to reduce the large network parameters of deep models.

REFERENCES

[1] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Trans. Affective Comput.*, to be published. doi: 10.1109/TAFFC.2017.2731763.

[2] X. Zhao and S. Zhang, "A review on facial expression recognition: Feature extraction and classification," *IETE Tech. Rev.*, vol. 33, no. 5, pp. 505–517, 2016.

[3] C. A. Corneanu, M. O. Simón, J. F. Cohn, and S. E. Guerrero, "Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 8, pp. 1548–1568, Aug. 2016.

[4] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 6, pp. 1113–1133, Jun. 2015.

[5] G. Muhammad, M. Alsulaiman, S. U. Amin, A. Ghoneim, and M. F. Alhamid, "A facial-expression monitoring system for improved healthcare in smart cities," *IEEE Access*, vol. 5, pp. 10871–10881, 2017.

[6] T. Wu, M. S. Bartlett, and J. R. Movellan, "Facial expression recognition using Gabor motion energy filters," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. Workshops*, San Francisco, CA, USA, Jun. 2010, pp. 42–47.

[7] G. Zhao and M. Pietikäinen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 29, no. 6, pp. 915–928, Jun. 2007.

[8] B. Jiang, M. F. Valstar, B. Martinez, and M. Pantic, "A dynamic appearance descriptor approach to facial actions temporal modeling," *IEEE Trans. Cybern.*, vol. 44, no. 2, pp. 161–174, Feb. 2014.

[9] T. Zhang, W. Zheng, Z. Cui, Y. Zong, J. Yan, and K. Yan, "A deep neural network-driven feature learning method for multi-view facial expression recognition," *IEEE Trans. Multimedia*, vol. 18, no. 12, pp. 2528–2536, Dec. 2016.

[10] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Dec. 2015, pp. 2983–2991.

[11] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Trans. Image Process.*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017.

[12] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3D convolutional neural networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops*, Jul. 2017, pp. 2278–2288.

[13] A. T. Lopes, E. de Aguiar, A. F. De Souza, and T. Oliveira-Santos, "Facial expression recognition with convolutional neural networks: Coping with few data and the training sample order," *Pattern Recognit.*, vol. 61, pp. 610–628, Jan. 2017.

[14] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Proc. Asian Conf. Comput. Vis. (ACCV)*, Singapore, 2014, pp. 143–157.

[15] S. Zhang, S. Zhang, T. Huang, W. Gao, and Q. Tian, "Learning affective features with a hybrid deep model for audio–visual emotion recognition," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 28, no. 10, pp. 3030–3043, Oct. 2018.

[16] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," in *Proc. ICLR*, San Diego, CA, USA, 2015, pp. 1–14.

[17] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *Science*, vol. 313, no. 5786, pp. 504–507, 2006.

[18] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Proc. Int. Conf. Neural Inf. Process. Syst.*, Montreal, QC, Canada, 2014, pp. 568–576.

[19] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, "Multimodal deep learning," in *Proc. 28th Int. Conf. Mach. Learn. (ICML)*, 2011, pp. 689–696.

[20] S. Zhalehpour, O. Onder, Z. Akhtar, and C. E. Erdem, "BAUM-1: A spontaneous audio-visual face database of affective and mental states," *IEEE Trans. Affective Comput.*, vol. 8, no. 3, pp. 300–313, Jul./Sep. 2016.

[21] Y. Wang, L. Guan, and A. N. Venetsanopoulos, "Kernel cross-modal factor analysis for information fusion with application to bimodal emotion recognition," *IEEE Trans. Multimedia*, vol. 14, no. 3, pp. 597–607, Jun. 2012.

[22] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *Proc. IEEE Int. Conf. Multimedia Expo (ICME)*, Amsterdam, The Netherlands, Jul. 2005, pp. 317–321.

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

[24] U. Tariq, J. Yang, and T. S. Huang, "Multi-view facial expression recognition analysis with generic sparse coding feature," in *Proc. Eur. Conf. Comput. Vis. (ECCV)*, 2012, pp. 578–588.

[25] U. Tariq *et al.*, "Emotion recognition from an ensemble of features," in *Proc. IEEE Int. Conf. Autom. Face Gesture Recognit. Workshops (FG)*, Santa Barbara, CA, USA, Mar. 2011, pp. 872–877.

[26] A. Klaser, M. Marszałek, and C. Schmid, "A spatio-temporal descriptor based on 3D-gradients," in *Proc. 19th Brit. Mach. Vis. Conf. (BMVC)*, vol. 275, 2008, pp. 1–10.

[27] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proc. 15th ACM Int. Conf. Multimedia (MM)*, Augsburg, Germany, 2007, pp. 357–360.

[28] M. Hayat, M. Bennamoun, and A. El-Sallam, "Evaluation of spatiotemporal detectors and descriptors for facial expression recognition," in *Proc. 5th Int. Conf. Hum. Syst. Interact. (HSI)*, Perth, WA, Australia, 2012, pp. 43–47.

[29] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets via universal manifold model for dynamic facial expression recognition," *IEEE Trans. Image Process.*, vol. 25, no. 12, pp. 5920–5932, Dec. 2016.

[30] X. Fan and T. Tjahjadi, "A dynamic framework based on local Zernike moment and motion history image for facial expression recognition," *Pattern Recognit.*, vol. 64, pp. 399–406, Apr. 2017.

[31] H. Yan, "Collaborative discriminative multi-metric learning for facial expression recognition in video," *Pattern Recognit.*, vol. 75, pp. 33–40, Mar. 2018.

[32] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "ImageNet classification with deep convolutional neural networks," in *Proc. Adv. Neural Inf. Process. Syst.*, 2012, pp. 1106–1114.

[33] C. Szegedy *et al.*, "Going deeper with convolutions," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 1–9.

[34] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Las Vegas, NV, USA, Jun. 2016, pp. 770–778.

[35] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proc. IEEE*, vol. 86, no. 11, pp. 2278–2324, Nov. 1998.

[36] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3D convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Santiago, Chile, Dec. 2015, pp. 4489–4497.

[37] G. Gkioxari and J. Malik, "Finding action tubes," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Boston, MA, USA, Jun. 2015, pp. 759–768.

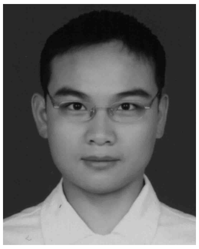
[38] P. Viola and M. J. Jones, "Robust real-time face detection," *Int. J. Comput. Vis.*, vol. 57, no. 2, pp. 137–154, 2004.

[39] G. E. Hinton, "Training products of experts by minimizing contrastive divergence," *Neural Comput.*, vol. 14, no. 8, pp. 1771–1800, 2002.

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

[41] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Proc. 12th IEEE Int. Conf. Autom. Face Gesture Recognit. (FG)*, Washington, DC, USA, May/Jun. 2017, pp. 790–795.

[42] N. El D. Elmadany, Y. He, and L. Guan, "Multiview emotion recognition via multi-set locality preserving canonical correlation analysis," in *Proc. IEEE Int. Symp. Circuits Syst. (ISCAS)*, Montreal, QC, Canada, May 2016, pp. 590–593.



SHIQING ZHANG received the Ph.D. degree from the School of Communication and Information Engineering, University of Electronic Science and Technology of China, in 2012. He held a Postdoctoral position with the School of Electronic Engineering and Computer Science, Peking University, Beijing, China, from 2015 to 2017. He is currently a Professor with the Institute of Intelligent Information Processing, Taizhou University, China. He has published over 30 papers in journals such as the IEEE Transactions on Multimedia and the IEEE Transactions on Circuits and Systems for Video Technology. His research interests include affective computing and pattern recognition.

XIANZHANG PAN received the B.S. and M.S. degrees in computer science from Lanzhou Jiaotong University. He is currently a Senior Engineer with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include image processing, affective computing, and pattern recognition.

YUELI CUI received the B.S. degree in electronics and communication engineering from Zhejiang University City College, Hangzhou, in 2006, and the M.S. degree in electronics and communication engineering from Hebei University, Baoding, in 2009. He is currently a Lecturer with the Department of Physics and Electronics Engineering, Taizhou University, China. His research interests include image processing and pattern recognition.

XIAOMING ZHAO received the B.S. degree in mathematics from Zhejiang Normal University, in 1990, and the M.S. degree in software engineering from Beihang University, in 2006. He is currently a Professor with the Institute of Intelligent Information Processing, Taizhou University, China. His research interests include image processing, machine learning, and pattern recognition.

LIMEI LIU received the Ph.D. degree from the School of Information Science and Engineering, Central South University, in 2011. She held a postdoctoral position with the Business School, Hunan University, Changsha, China, from 2013 to 2017. She is currently a Professor with the Institute of Big Data and Internet Innovation, Hunan University of Commerce, China. Her research interests include machine learning and pattern recognition.