SEMINAR REPORT

ON

LEARNING AFFECTIVE VIDEO FEATURES FOR FACIAL EXPRESSION RECOGNITION

Submitted by

SREEKANTH PAI

(KSD18CS082)

To the APJ Abdul Kalam Technological University in partial fulfillment of the requirements for the award of the degree of

BACHELOR OF TECHNOLOGY

in

COMPUTER SCIENCE AND ENGINEERING

DEPARTMENT OF COMPUTER SCIENCE ENGINEERING

LBS COLLEGE OF ENGINEERING

KASARAGOD-671542, KERALA

JANUARY 2021

DECLARATION

"I hereby declare that this submission is my own work and that, to the best of my knowledge and belief, it contains no material previously published or written by another person nor material which has been accepted for the award of any other degree or diploma of this Institute or other Institute of higher learning, except where due acknowledgement has been made in the text."

PLACE: KASARAGOD	NAME: SREEKANTH PAI
DATE: 12-01-2022	REG NO: KSD18CS082

CERTIFICATE

This is to certify that the seminar report entitled **"LEARNING AFFECTIVE VIDEO FEATURES FOR FACIAL EXPRESSION RECOGNITION"** *submitted by* **SREEKANTH PAI** *on* **12/01/2022** *to the APJ Abdul Kalam Technological University in partial fulfilment of the requirements for the award of the degree of* **Bachelor of Technology in Computer Science Engineering** *is a bonafide record of the work done by him under my supervision and guidance*.

SEMINAR COORDINATOR
(GUIDE)
SARITH DIVAKAR M
ASSISTANT PROFESSOR
CSE DEPARTMENT

HEAD OF DEPARTMENT
SMITHAMOL MB
ASSOCIATE PROFESSOR
CSE DEPARTMENT

Place: Kasaragod

Date: 12/01/2022

ACKNOWLEDGEMENT

It is a really a momentous opportunity and privilege to express my deep sense of gratitude to all those who have helped me to accomplish this task. I express my humble to God Almighty for his incessant blessing on me during this seminar. I am indebted to God almighty for being the guiding light throughout this work and for helping me to complete the same within the stipulated time

The efforts that I have taken for presenting the seminar would not have been possible without the kind support and help of many individuals. I would like to extend my sincere thanks to all of them.

I sincerely thank our principal **Dr. MOHAMMED SHEKOOR T** for providing me facilities to do my seminar presentation. I express my sincere gratitude to head of the department **Dr. SMITHAMOL M B** and I also express heartiest gratitude to seminar coordinator and my guide **Mr. SARITH DIVAKAR M** for his valuable advice and guidance. I would always oblige for the helping hands of all other staff members of the department and all my friends and well-wishers who directly or indirectly contributed to this venture.

SREEKANTH PAI

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.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 proposed method, outperforming the state-of-the-arts.

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

CONTENTS

LIST OF FIGURES

LIST OF ABBREVIATIONS

CNN	: Convolutional Neural Network
FER	: Facial Expression Recognition
DBN	: Deep Belief Network

1. 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, humancomputer interaction and healthcare . 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 . 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 , Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) or Local Phase Quantization from TOP (LPQ-TOP) . However, these hand-designed feature representations are low-level to discriminate dynamic facial expressions. Recently, the deep neural network driven feature learning representations from data may achieve better performance without requiring domain expertise.

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. Hybrid deep learning model contains three deep models. The first two deep models are deep Convolutional Neural Networks (CNNs) , 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) 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 . Nevertheless, 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 . 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 , the RML database , and the MMI database, 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) 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, a deep DBN model is employed 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 this method outperforms the-state-of-the-arts.

2. BACKGROUND

CONVOLUTIONAL NUERAL NETWORK(CNN)

A convolutional neural network (CNN) is a type of artificial neural network used in image recognition and processing that is specifically designed to process pixel data.CNNs are powerful image processing, artificial intelligence (AI) that use deep learning to perform both generative and descriptive tasks, often using machine vison that includes image and video recognition, along with recommender systems and natural language processing (NLP).

A neural network is a system of hardware and/or software patterned after the operation of neurons in the human brain. Traditional neural networks are not ideal for image processing and must be fed images in reduced-resolution pieces. CNN have their “neurons” arranged more like those of the frontal lobe, the area responsible for processing visual stimuli in humans and other animals. The layers of neurons are arranged in such a way as to cover the entire visual field avoiding the piecemeal image processing problem of traditional neural networks.

A CNN uses a system much like a multilayer perceptron that has been designed for reduced processing requirements. The layers of a CNN consist of an input layer, an output layer and a hidden layer that includes multiple convolutional layers, pooling layers, fully connected layers and normalization layers. The removal of limitations and increase in efficiency for image processing results in a system that is far more effective, simpler to trains limited for image processing and natural language processing.

DEEP BELIEF NETWORK(DBN)

Deep Belief Networks (DBNs) is the technique of stacking many individual unsupervised networks that use each network’s hidden layer as the input for the next layer. Usually, a “stack” of restricted Boltzman machines (RBMs) or autoencoders are employed in this role. The ultimate goal is to create a faster unsupervised training procedure that relies on contrastive divergence for each sub-network. The lowest visible layer is called the training set. From there, each layer can communicate with the previous and subsequent layers. However, the nodes of any particular layer cannot communicate laterally with each other.

In supervised learning, this stack usually ends with a final classification layer and in unsupervised learning it often ends with an input for cluster analysis. Except for the first and last layers, each level in a DBN serves a dual role function: it's the hidden layer for the nodes that came before and the visible (output) layer for the nodes that come next

3. EXISTING SYSTEM

A. HAND-DESIGNED FEATURE-BASED METHOD

For facial feature representation in static images, a variety of local image descriptors, including Local Binary Pattern (LBP), Histogram of Oriented Gradient (HOG) , and Scale Invariant Feature Transform (SIFT) 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 , LPQ-TOP , 3D-HOG , 3D-SIFT, respectively. Hayat et al 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 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. present an expressionletbased 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 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.

Recently, some efforts have been conducted to develop more powerful spatio-temporal feature extraction methods for FER. For instance, Liu et al. present an expressionletbased spatio-temporal manifold descriptor which shows the superiority over traditional methods on FER tasks. Fan and Tjahjadi provide a spatio-temporal feature based on local Zernike moment and motion history image for dynamic FER. Yan 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, 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 into a deep multilayered 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 a video sequence. To solve this problem, the recentlydeveloped 3D-CNNs 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 . In addition, a variant of 3D-CNNs is 3DCNN-DAP 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. 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 , may present a cue. However, the used shallow fusion method in 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, a deep fusion network built with a deep DBN model is designed to jointly learn the outputs of two-stream CNNs. Then apply this hybrid deep learning model for FER in video sequences. Experiment results on three video-based facial expression databases demonstrate the advantages of proposed method.

4. METHODOLOGY

Figure 1 shows the framework of proposed hybrid deep learning model. As depicted in Fig.1, This 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, this 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, present the details about abovementioned four steps of our method.

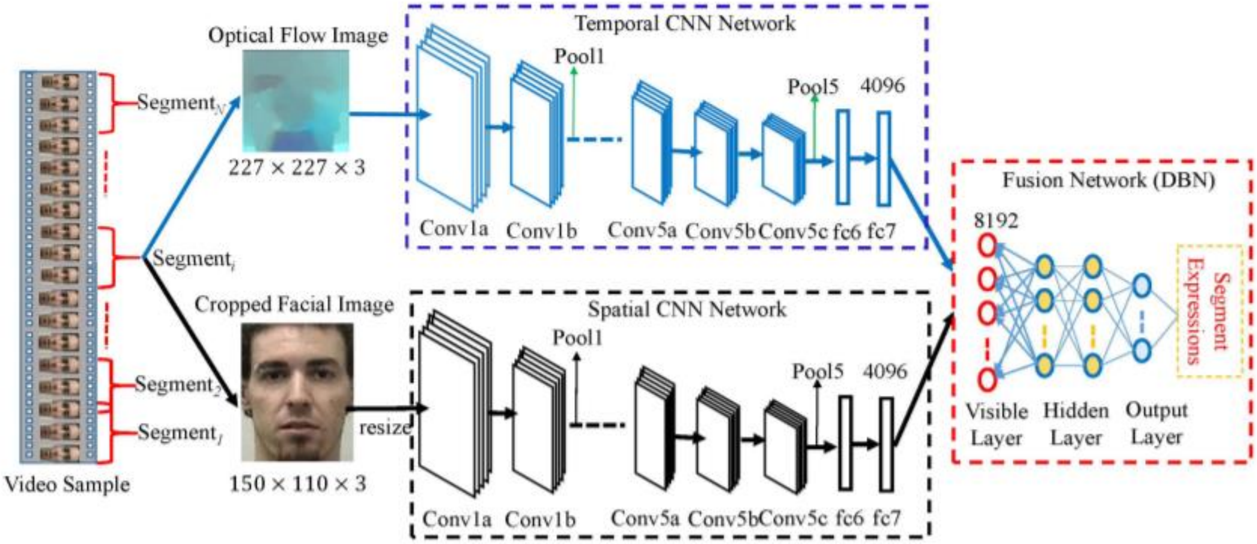

Fig 1. The framework of hybrid deep learning network for facial expression recognition in video sequences

A. GENERATION OF CNN INPUTS

Since CNNs require a fixed size of input data, each video sample is divided 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 ,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, each video sample is divided into a fixed-length segment with L = 16. To this end, when L > 16 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, make sure that each divided segment has a length of L = 16.

1) INPUTS OF TEMPORAL CNNs

To produce suitable inputs of temporal CNNs, extract optical flow images between consecutive frames in a video sequence. Optical flow images represent the displacement changes of corresponding positions between consecutive frames. The transformed flow maps are conserved as an optical flow image containing three channels, which corresponds to motion d˜ x , d˜ y and the optical flow magnitude. It produces 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, employ a cropped facial image of 150 × 110 × 3 for each frame in a video segment, as in . In detail, a robust real-time face detector 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, resize the cropped facial image into 227 × 227 × 3 as inputs of spatial CNNs. Note that discard the first frame in a video segment L = 16, and employ the remaining 15 frames as inputs of spatial CNNs. In this case, 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 , which consists of five convolution layers (Conv1a-Conv1b, Conv2aConv2b, Conv3a-Conv3b-Conv3c- , · · · , Conv5a, Conv5bConv5c), five max-pooling layers (Pool1, Pool2, · · · , 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 on target video-based facial expression data. In particular, we firstly copy the existing VGG16 parameters pre-trained on largescale ImageNet data to initialize the temporal CNN network and the spatial CNN network, respectively. Then, replace the fc8 layer in VGG16 with a new class label vector corresponding to six facial expression categories used in the experiments. Ultimately, individually retrain these two CNN streams by using the standard back propagation strategy. Specially, use the back propagation technique to solve the following minimizing problem so as to update the CNN network parameters.

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 , 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.

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

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

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 segmentlevel features of DBNs, 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.

5. EXPERIMENTATION

To verify the performance of 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 , the RML database and the MMI database .

A. DATASETS

1) BAUM-1s

The original BAUM-1 is a newly-developed spontaneous audio-visual face database of affective and mental states . 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.

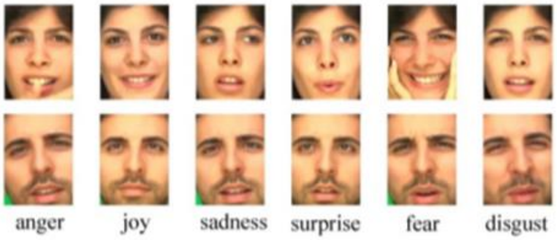

Fig 2. Some examples of cropped facial expression images from the BAUM-1s dataset.

2) RML

The RML database 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.

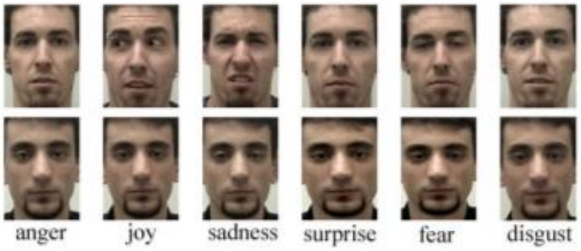

Fig 3. Some examples of cropped facial expression images from the RML dataset.

3) MMI

The MMI database 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.

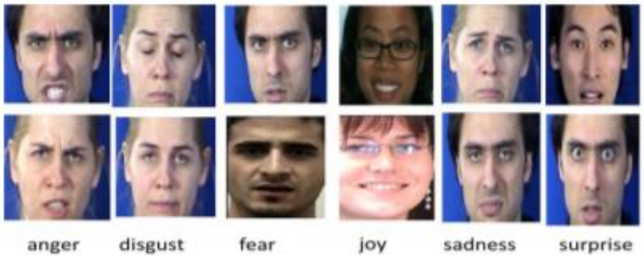

Fig 4. Some examples of cropped facial expression images from the MML dataset

6. 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 proposed method in two stages. (1)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 proposed method.For instance, it is challenging to develop a real-time FER system based on 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.

7. REFERENCES

S. Zhang, X. Pan, Y. Cui, X. Zhao and L. Liu, "Learning Affective Video Features for Facial Expression Recognition via Hybrid Deep Learning," in IEEE Access, vol. 7, pp. 32297-32304, 2019, doi: 10.1109/ACCESS.2019.2901521.

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.

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.

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.

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.

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

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.

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.

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.

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.

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.

K. Simonyan and A. Zisserman, ‘‘Very deep convolutional networks for large-scale image recognition,’’ in Proc. ICLR, San Diego, CA, USA, 2015.