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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

Deep Facial Expression Recognition: A Survey

Shan Li and Weihong Deng*, *Member, IEEE*

Abstract—With the transition of facial expression recognition (FER) from laboratory-controlled to challenging in-the-wild conditions and the recent success of deep learning techniques in various fields, deep neural networks have increasingly been leveraged to learn discriminative representations for automatic FER. Recent deep FER systems generally focus on two important issues: overfitting caused by a lack of sufficient training data and expression-unrelated variations, such as illumination, head pose and identity bias. In this survey, we provide a comprehensive review of deep FER, including datasets and algorithms that provide insights into these intrinsic problems. First, we introduce the available datasets that are widely used in the literature and provide accepted data selection and evaluation principles for these datasets. We then describe the standard pipeline of a deep FER system with the related background knowledge and suggestions for applicable implementations for each stage. For the state-of-the-art in deep FER, we introduce existing novel deep neural networks and related training strategies that are designed for FER based on both static images and dynamic image sequences and discuss their advantages and limitations. Competitive performances and experimental comparisons on widely used benchmarks are also summarized. We then extend our survey to additional related issues and application scenarios. Finally, we review the remaining challenges and corresponding opportunities in this field as well as future directions for the design of robust deep FER systems.

Index Terms—Facial Expression Recognition, Facial Expression Datasets, Affect, Deep Learning, Survey.

1 Introduction

Facial expression is one of the most powerful, natural and universal signals for human beings to convey their emotional states and intentions [1], [2]. Numerous studies have been conducted on automatic facial expression analysis because of its practical importance in sociable robots, medical treatment, driver fatigue surveillance, and many other human-computer interaction systems. In the field of computer vision and machine learning, various facial expression recognition (FER) systems have been explored to encode expression information from facial representations. As early as the twentieth century, Ekman and Friesen [3] defined six basic emotions based on a cross-cultural study [4], which indicated that humans perceive certain basic emotions in the same way regardless of culture. These prototypical facial expressions are anger, disgust, fear, happiness, sadness, and surprise. Contempt was subsequently added as one of the basic emotions [5]. Recently, advanced research on neuroscience and psychology argued that the model of six basic emotions is culture-specific and not universal [6].

Although the affect model based on basic emotions is limited in the ability to represent the complexity and subtlety of our daily affective displays [7], [8], [9], and other emotion description models, such as the facial action coding system (FACS) [10] and the continuous model using affect dimensions [11], are considered to represent a wider range of emotions. The categorical model that describes emotions in terms of discrete basic emotions is still the most popular perspective for FER due to its pioneering investigations along with the direct and intuitive definition of facial expressions. In this survey, we limit our discussion on FER based on the categorical model.

FER systems can be divided into two main categories according to the feature representations: static image FER and dynamic sequence FER. In *static-based methods* [12], [13], [14], the feature representation is encoded with only spatial information from the current single image, whereas *dynamic-based methods* [15], [16], [17] consider the temporal relation among contiguous frames in the input facial expression sequence. Based on these two vision-based methods, other modalities, such as audio and physiological channels, have also been used in *multimodal systems* [18] to assist in the recognition of expression. Although pure expression recognition based on visible face images can achieve promising results, incorporating other models into a high-level framework can provide complementary information and further enhance robustness.

The majority of the traditional methods have used handcrafted features or shallow learning (e.g., local binary patterns (LBP) [12], LBP on three orthogonal planes (LBP-TOP) [15], non-negative matrix factorization (NMF) [19] and sparse learning [20]) for FER. However, since 2013, emotion recognition competitions such as FER2013 [21] and Emotion Recognition in the Wild (EmotiW) [22], [23] have collected relatively sufficient training data from challenging real-world scenarios, which implicitly promote the transition of FER from lab-controlled to in-the-wild settings. Additionally, due to the dramatically increased chip processing abilities (e.g., GPU units) and well-designed network architecture, studies in various fields have begun to transfer to deep learning methods, which have achieved state-of-the-art recognition accuracy and exceeded previous results by a large margin (e.g., [24], [25], [26], [27]). Similarly, given the more effective training data of facial expressions, deep learning techniques have increasingly been implemented to handle the challenging factors for emotion recognition in the wild. Figure 1 illustrates this evolution of FER in terms of algorithms and datasets.

Exhaustive surveys on automatic expression analysis have been published in recent years [7], [8], [28], [29]. These surveys have established a set of standard algorithmic pipelines for FER. However, they focus on traditional methods, and deep learning has rarely been reviewed. Very recently, deep learning for human

• *The authors are with the Pattern Recognition and Intelligent System Laboratory, School of Information and Communication Engineering, Beijing University of Posts and Telecommunications, Beijing, 100876, China. E-mail:{ls1995, whdeng}@bupt.edu.cn.*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

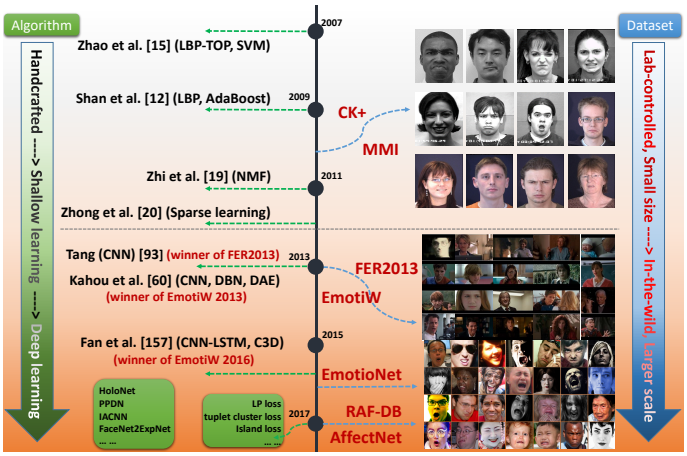


Fig. 1. The evolution of facial expression recognition in terms of datasets and methods.

affect recognition was surveyed in [30], which reviewed the development of deep affect recognition from 2010 to 2017 and focused on the fusion of audiovisual and physiological sensors. In this paper, we conduct more specific and detailed research on deep learning for both static and dynamic FER tasks until 2019. We aim to give a newcomer to this field an overview of the systematic framework and prime skills for deep FER.

Despite the powerful feature learning ability of deep learning, problems remain when applied to FER. First, deep neural networks require a large quantity of training data to avoid overfitting. However, the existing facial expression databases are not sufficient to train the well-known neural network with deep architecture that achieved the most promising results in object recognition tasks. Additionally, high intersubject variations exist due to different personal attributes, such as age, gender, ethnic backgrounds and level of expressiveness [31]. In addition to subject identity bias, variations in pose, illumination and occlusions are common in unconstrained facial expression scenarios. These disturbances are nonlinearly confounded with facial expressions and therefore strengthen the requirement of deep networks to address the large intraclass variability and to learn effective expression-specific representations.

In this paper, we introduce recent advances in research on solving the above problems for deep FER. We examine the state-of-the-art results that have not been reviewed in previous survey papers. The rest of this paper is organized as follows. Frequently used expression databases are introduced in Section 2. Section 3 identifies three main steps required in a deep FER system and describes the related background. Section 4 provides a detailed review of novel neural network architectures and special network training tricks designed for FER based on static images and dynamic image sequences. We then cover additional related issues and other practical scenarios in Section 5. Section 6 discusses some of the challenges and opportunities in this field and identifies potential future directions.

2 Facial Expression Databases

Having sufficient labeled training data that include as many variations of the populations and environments as possible is important for the design of a deep expression recognition system. In this section, we discuss publicly available databases that contain basic expressions and that are widely used in our reviewed papers for deep learning algorithm evaluation. We also introduce newly released databases that contain a large number of affective images collected from the real world to benefit the training of deep neural networks. Table 1 provides an overview of these datasets, including the main reference, number of subjects, number of images or video samples, collection environment, expression distribution and additional information. Figure 2 exhibits facial expression images collected from laboratory and real-world conditions.

Fig. 2. Sample images with seven basic emotions collected from different environments (lab-controlled v.s. real world). Photograph (one per row): CK+ [32], Oulu-CASIA [33]; RAF-DB [34], EmotiW 2017 [23].

CK+ [32]: The Extended Cohn–Kanade (CK+) database is the most extensively used laboratory-controlled database for evaluating FER systems. CK+ contains sequences that show a shift from neutral expression to peak expression. For evaluation, the most common data selection method is to extract the last one to three frames with peak formation and the first frame of each sequence. Then, the subjects are divided into n groups for person-independent n-fold cross-validation experiments, where commonly selected values of n are 5, 8 and 10.

MMI [35]: The MMI database is also laboratory controlled. In contrast to CK+, sequences in MMI are onset-apex-offset labeled, i.e., the sequence begins with a neutral expression and reaches a peak near the middle before returning to the neutral expression. For experiments, the most common method is to choose the first frame (neutral face) and the three peak frames in each frontal sequence to conduct person-independent 10-fold cross-validation.

Oulu-CASIA [33]: The Oulu-CASIA database includes 2,880 image sequences collected from 80 subjects. Each of the videos is captured with one of two imaging systems, i.e., near-infrared (NIR) or visible light (VIS), under three different illumination conditions. Similar to CK+, the first frame is neutral, and the last frame has the peak expression. Typically, only the last three peak frames and the first frame (neutral face) from the 480 videos collected by the VIS system under normal indoor illumination are employed for 10-fold cross-validation experiments.

JAFFE [36]: The Japanese Female Facial Expression (JAFFE) database contains 213 samples of posed expressions from 10 Japanese females. Each person has 3˜4 images with each of six basic facial expressions and one image with a neutral expression. Typically, all the images are used for the leave-one-subject-out experiment.

FER2013 [21]: FER2013 is a large-scale and unconstrained database collected automatically by the Google image search API. All images were registered and resized to 48*48 pixels after

Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

TABLE 1
An overview of the facial expression datasets. P = posed; S = spontaneous; Condit. = Collection condition; Elicit. = Elicitation method.

Database	Samples	Subject	Condit.	Elicit.	Expression distribution	Access
CK+ [32]	593 image sequences	123	Lab	P & S	seven basic expressions plus contempt	http://www.consortium.ri.cmu.edu/ckagree/
MMI [35]	740 images and 2,900 videos	25	Lab	P	seven basic expressions	https://mmifacedb.eu/
JAFFE [36]	213 images	10	Lab	P	seven basic expressions	http://www.kasrl.org/jaffe.html
TFD [37]	112,234 images	N/A	Lab	P	seven basic expressions	josh@mplab.ucsd.edu
FER-2013 [21]	35,887 images	N/A	Web	P & S	seven basic expressions	https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge
AFEW 7.0 [23]	1,809 videos	N/A	Movie	P & S	seven basic expressions	https://sites.google.com/site/emotiwchallenge/
SFEW 2.0 [22]	1,766 images	N/A	Movie	P & S	seven basic expressions	https://cs.anu.edu.au/few/emotiw2015.html
Multi-PIE [38]	755,370 images	337	Lab	P	Smile, surprised, squint, disgust, scream and neutral	http://www.flintbox.com/public/project/4742/
BU-3DFE [39]	2,500 3D images	100	Lab	P	seven basic expressions	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
BU-4DFE [40]	606 3D sequences	101	Lab	P	seven basic expressions	http://www.cs.binghamton.edu/~lijun/Research/3DFE/3DFE_Analysis.html
Oulu-CASIA [33]	2,880 image sequences	80	Lab	P	six basic expressions without neutral	http://www.cse.oulu.fi/CMV/Downloads/Oulu-CASIA
RaFD [41]	1,608 images	67	Lab	P	seven basic expressions plus contempt	http://www.socsci.ru.nl:8180/RaFD2/RaFD
KDEF [42]	4,900 images	70	Lab	P	seven basic expressions	http://www.emotionlab.se/kdef/
EmotioNet [43]	1,000,000 images	N/A	Internet	P & S	23 basic expressions or compound expressions	http://cbcsl.ece.ohio-state.edu/dbform_emotionet.html
RAF-DB [34], [44]	29672 images	N/A	Internet	P & S	seven basic expressions and twelve compound expressions	http://www.whdeng.cn/RAF/model1.html
AffectNet [45]	450,000 images (labeled)	N/A	Internet	P & S	seven basic expressions	http://mohammadmahoor.com/databases-codes/
ExpW [46]	91,793 images	N/A	Internet	P & S	seven basic expressions	http://mmlab.ie.cuhk.edu.hk/projects/socialrelation/index.html
4DFAB [47]	1.8 million 3D faces	180	Lab	P & S	seven basic expression	N/A

rejecting incorrectly labeled frames and adjusting the cropped region. FER2013 contains 28,709 training images, 3,589 validation images and 3,589 test images with seven expression labels.

AFEW [48] and SFEW [49]: The Acted Facial Expressions in the Wild (AFEW) database contains video clips collected from different movies with spontaneous expressions, various head poses, occlusions and illuminations. AFEW is a temporal and multimodal database that provides vastly different environmental conditions in both audio and video. The AFEW is independently divided into three data partitions in terms of subject and movie/TV source, which ensures data in the three sets belong to mutually exclusive movies and actors. The Static Facial Expressions in the Wild (SFEW) was created by selecting static frames from the AFEW database. The most commonly used version, SFEW 2.0, has been divided into three sets: Train, Val and Test. The expression labels of the training and validation sets are publicly available, whereas those of the testing set are held back by the challenge organizer.

Multi-PIE [38]: The CMU Multi-PIE database contains 755,370 images from 337 subjects under 15 viewpoints and 19 illumination conditions in up to four recording sessions. Each facial image is labeled with one of six expressions. This dataset is typically used for multiview facial expression analysis.

BU-3DFE [39] and BU-4DFE [40]: The Binghamton University 3D Facial Expression (BU-3DFE) database contains 606 facial expression sequences captured from 100 people. For each subject, six facial expressions are elicited in various manners with multiple intensities. Similar to Multi-PIE, this dataset is typically used for multiview 3D facial expression analysis. To analyze the facial behavior from a static 3D space to a dynamic 3D space, BU-4DFE was constructed, which contains 606 3D facial expression sequences with a total of approximately 60,600 frame models.

EmotioNet [43]: EmotioNet is a large-scale database with one million facial expression images collected from the Internet. A total of 950,000 images were annotated by the automatic action unit (AU) detection model in [43], and the remaining 25,000 images were manually annotated with 11 AUs. The second track of the EmotioNet Challenge [50] provides six basic expressions and ten compound expressions [51], and 2,478 images with expression labels are available.

RAF-DB [34], [44]: The Real-world Affective Face Database (RAF-DB) is a real-world database that contains 29,672 highly diverse facial images downloaded from the Internet. With manually crowd-sourced annotation and reliable estimation, seven basic and eleven compound emotion labels are provided for the samples. Specifically, 15,339 images from the basic emotion set are divided into two groups (12,271 training samples and 3,068 testing samples) for evaluation.

AffectNet [45]: AffectNet contains more than one million images from the Internet that were obtained by querying different search engines using emotion-related tags. It is by far the largest database that provides facial expressions in two different emotion models (categorical model and dimensional model), of which 450,000 images have manually annotated labels for eight basic expressions.

ExpW [46]: The Expression in-the-Wild Database (ExpW) contains 91,793 faces downloaded using Google image search. Each of the face images was manually annotated as one of the seven basic expression categories. Nonface images were removed in the annotation process.

4DFAB [47]: 4DFAB is a large-scale database with over 1,800,000 high-resolution 3D faces, which has records of 180 subjects captured in four different sessions spanning a five-year period. It contains 4D dynamic videos of subjects displaying both spontaneous and posed facial behaviors of six basic expressions.

3 Deep Facial Expression Recognition

In this section, we describe the three main steps that are common in automatic deep FER, i.e., preprocessing, deep feature learning and deep feature classification. We briefly summarize the widely used algorithms for each step and recommend the existing state-of-the-art best practice implementations according to the referenced papers.

3.1 Preprocessing

Variations that are irrelevant to facial expressions, such as different backgrounds, illuminations and head poses, are fairly common in

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

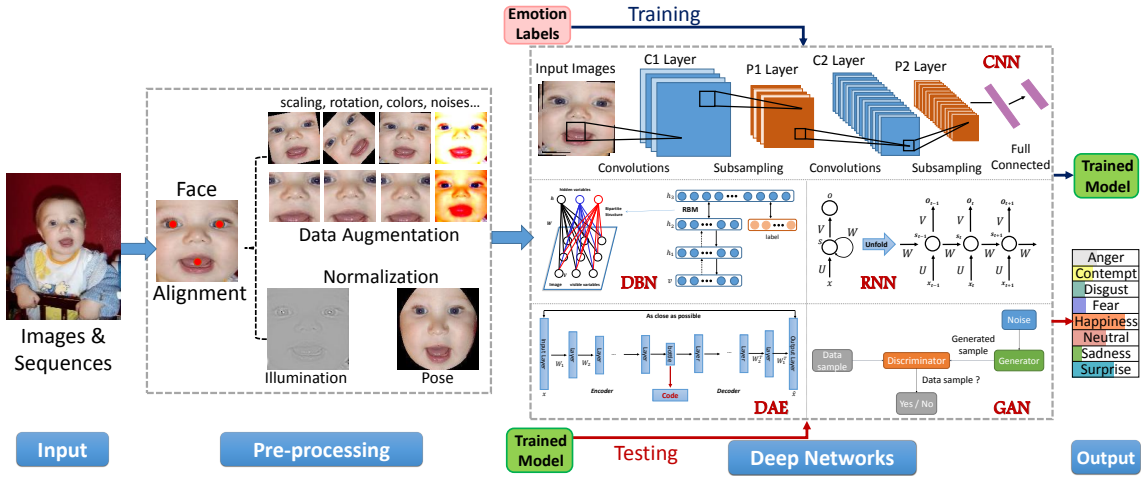


Fig. 3. The general pipeline of deep facial expression recognition systems.

unconstrained scenarios. Therefore, before training the deep neural network to learn meaningful features, preprocessing is usually required to align and normalize the visual semantic information conveyed by the face.

3.1.1 Face alignment

We list some well-known approaches and publicly available face alignment implementations that are widely used in deep FER. Given a series of training data, the first step is to detect the face and then to remove background and nonface areas. The Viola-Jones (V&J) face detector [52] is a classic and widely employed implementation for face detection, which is robust and computationally simple for detecting near-frontal faces.

Although face detection is the only indispensable procedure to enable feature learning, further face alignment using the coordinates of localized landmarks can substantially enhance the FER performance [14]. This step is crucial because it can reduce the variation in facial size and in-plane rotation. Table 2 investigates facial landmark detection algorithms widely used in deep FER and compares them in terms of efficiency and performance. In general, cascaded regression combined with deep networks has become the most popular and state-of-the-art method for face alignment due to its high speed and accuracy.

In contrast to using only one detector for face alignment, some methods proposed combining multiple detectors for better landmark estimation when processing faces in challenging unconstrained environments. Yu et al. [53] concatenated three different facial landmark detectors to complement each other. Kim et al. [54] considered different inputs (original image and histogram equalized image) and different face detection models (V&J [52] and MoT [55]) and chose the landmark result with the highest confidence predicted by the Intraface [56] for face alignment.

3.1.2 Data augmentation

Deep neural networks require sufficient training data to ensure generalizability to a given recognition task. However, most publicly available databases for FER do not have a sufficient quantity of images for training. Therefore, data augmentation is a vital step for deep FER. Data augmentation approaches can be divided into two types: on-the-fly data augmentation and offline data augmentation.

Usually, on-the-fly data augmentation is embedded in deep learning toolkits to alleviate overfitting. During the training step,

TABLE 2
Summary of different types of face alignment detectors that are widely used in deep FER models.

type		# points	real-time	speed	performance	used in
Holistic	AAM [57]	68	✗	fair	poor generalization	[58], [59]
Part-based	MoT [55]	39/68	✗	slow/ fast	good	[60], [61], [62]
	DRMF [63]	66	✗			[64], [65]
Cascaded regression	SDM [66]	49	✓	fast/ very fast	good/ very good	[14], [16], [67], [68]
	3000 fps [69]	68	✓			[59]
	Incremental [70]	49	✓			[71]
Deep learning	Cascaded CNN [72]	5	✓	fast	good/ very good	[73]
	MTCNN [74]	5	✓			[75], [76], [77]

the input samples are randomly cropped from the center and four corners of the image and then flipped horizontally, which can result in a dataset that is ten times larger than the original training data. Two common prediction modes are adopted during testing: only the center patch of the face is used for prediction (e.g., [65], [78]), or the prediction value is averaged over all ten crops (e.g., [54], [79]).

In addition to elementary on-the-fly data augmentation, various offline data augmentation operations have been designed to further expand data on both size and diversity. The most frequently used operations include random perturbations and transforms, e.g., rotation, shifting, skew, scaling, noise, contrast and color jittering. Combinations of multiple operations [53], [80] can generate more unseen training samples and make the network more robust to deviated and rotated faces. Furthermore, deep learning-based technology can be applied for data augmentation. For example, a synthetic data generation system with a 3D convolutional neural network (CNN) was created in [81] to confidentially create faces with different levels of saturation in expression. The generative adversarial network (GAN) [82] can also be applied to augment data by generating diverse appearances varying in poses and expressions (see Section 4.3.6).

3.1.3 Face normalization

Variations in illumination and head poses can introduce large changes in images and hence impair the FER performance.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

Therefore, we introduce two typical face normalization methods to ameliorate these variations: illumination normalization and pose normalization.

Illumination normalization: Illumination and contrast can vary in different images even from the same person with the same expression, especially in unconstrained environments, which can result in large intraclass variances. Various algorithms, such as isotropic diffusion (IS)-based normalization, discrete cosine transform (DCT)-based normalization, difference of Gaussian (DoG) and homomorphic filtering-based normalization, can be used for illumination normalization [64], [83]. Moreover, related studies have shown that histogram equalization combined with illumination normalization results in better face recognition performance than that achieved using illumination normalization alone. Many studies in the literature of deep FER (e.g., [53], [84], [85], [86]) have employed histogram equalization to increase the global contrast of images for preprocessing. This method is effective when the brightness of the background and foreground are similar. However, directly applying histogram equalization may overemphasize local contrast. To solve this problem, [87] proposed a weighted summation approach to combine histogram equalization and linear mapping.

Pose normalization: Pose variation is another common and intractable problem in unconstrained settings. Some studies have employed pose normalization techniques to yield frontal facial views for FER (e.g., [88], [89]), among which the most popular was proposed by Hassner et al. [90]. Specifically, after localizing facial landmarks, a 3D texture model generic to all faces is generated to estimate visible facial components. Then, the initial frontalized face is synthesized by backprojecting each input face image to the reference coordinate system. Alternatively, Sagonas et al. [91] proposed a statistical model that simultaneously localizes landmarks and converts facial poses using only frontal faces. Very recently, a series of GAN-based deep models were proposed for frontal view synthesis and reported promising performances.

3.2 Deep networks for feature learning

Deep learning has recently become a popular research topic and has achieved state-of-the-art performances for a variety of applications [92]. Deep learning attempts to capture high-level abstractions through hierarchical architectures of multiple nonlinear transformations and representations. In this section, we briefly introduce some deep learning techniques that have been applied for FER. The traditional architectures of these deep neural networks are shown in Fig. 3. Due to space limitations, we introduce the principles and applications of these networks, including convolutional neural networks, deep belief networks, deep autoencoders, recurrent neural networks and generative adversarial networks, in the supplementary material (See Section 1) in detail.

3.3 Facial expression classification

After learning the deep features, the final step of FER is to classify the given image into one of the basic emotion categories.

Unlike traditional methods, where the feature extraction step and the feature classification step are independent, deep networks can perform FER in an end-to-end way. Specifically, a loss layer is added to the end of the network to regulate the backpropagation error; then, the prediction probability of each sample can be directly output by the network. In CNN, softmax loss is the most commonly used function that minimizes the cross-entropy between the estimated class probabilities and the ground-truth distribution. Alternatively, [93] demonstrated the benefit of using a linear support vector machine (SVM) for end-to-end training, which minimizes margin-based loss instead of cross-entropy. Likewise, [94] investigated the adaptation of deep neural forests (NFs) [95], which replaces the softmax loss with NFs and achieved competitive results for FER.

In addition to the end-to-end learning method, another alternative is to employ the deep neural network (particularly a CNN) as a feature extraction tool and then apply additional independent classifiers, such as support vector machine or random forest, to the extracted representations [96], [97]. Furthermore, [98], [99] showed that the covariance descriptors computed on DCNN features and classification with Gaussian kernels on the symmetric positive definition (SPD) manifold are more efficient than the standard classification with the softmax layer.

4 The State-of-the-Art

In this section, we first introduce the specific pretraining and fine-tuning skills and diverse network inputs that are designed for FER. Then, we divide the works presented in the literature into two main groups depending on the type of data: deep FER networks for static images and deep FER networks for dynamic image sequences and discuss different network types that have been proposed in these groups. Moreover, we provide an overview of the current deep FER systems with respect to network architecture and performance.

4.1 Pretraining and fine-tuning

As mentioned before, direct training of deep networks on relatively small facial expression datasets is prone to overfitting. To mitigate this problem, many studies used additional task-oriented data to pretrain their self-built networks from scratch or fine-tuned on well-known pretrained models (e.g., AlexNet [24], VGG [25], VGG-face [114] and GoogleNet [26]). Kahou et al. [60], [115] indicated that the use of additional data can help to obtain models with high capacity without overfitting, thereby enhancing the FER performance.

To select appropriate auxiliary data, large-scale face recognition (FR) datasets or relatively large FER datasets are suitable. Kaya et al. [116] suggested that VGG-Face, which was trained for FR, overwhelmed ImageNet, which was developed for objected recognition. Another interesting result observed by Knyazev et al. [117] is that pretraining on a larger FR dataset can positively affect the expression recognition performance, and further fine-tuning with additional FER datasets can help improve the performance.

Instead of directly using the pretrained or fine-tuned models to extract features on the target dataset, a multistage fine-tuning strategy [67] can achieve better performance; after the first-stage fine-tuning using FER2013 on pretrained models, a second-stage fine-tuning using the training set of the target database (EmotiW) is employed to refine the models to adapt to a more specific dataset (i.e., the target dataset).

Although pretraining and fine-tuning on external FR data can indirectly avoid the problem of small training data, the networks are trained separately from the FER, and the face-dominated information remains in the learned features, which may weaken the

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

TABLE 3
Performance summary of representative methods for static-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Preprocessing = face detection & data augmentation & face normalization; IN = illumination normalization; $\mathcal{NE}$ = network ensemble; $\mathcal{CN}$ = cascaded network; $\mathcal{MN}$ = multitask network; LOSO = leave-one-subject-out.

Datasets	Method	Network type		Network size		Preprocessing			Data selection	Data group	Additional classifier	Performance[1](%)
CK+	Ouellet 14 [100]	CNN (AlexNet)		-	-	V&J	-	-	the last frame	LOSO	SVM	7 classes[‡] (94.4)
	Li et al. 15 [83]	RBM		4	-	V&J	-	IN			✗	6 classes: 96.8
	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	the last three frames and the first frame	8 folds	AdaBoost	6 classes: 96.7
	Liu et al. 13 [101]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	8 classes: 92.05 (87.67)
	Liu et al. 15 [102]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes[‡] 93.70
	Khorrami et al. 15 [103]	zero-bias CNN		4	7m	✓	✓	-		10 folds	✗	6 classes: 95.7; 8 classes: 95.1
	Ding et al. 17 [68]	CNN	fine-tune	8	11m	IntraFace	✓	-		10 folds	✗	6 classes: (98.6); 8 classes: (96.8)
	Zeng et al. 18 [58]	DAE (DSAE)		3	-	AAM	-	-	the last four frames and the first frame	LOSO	✗	7 classes[‡] 95.79 (93.78) 8 classes: 89.84 (86.82)
	Cai et al. 17 [104]	CNN	loss layer	6	-	DRMF	✓	IN	the last three frames	10 folds	✗	7 classes[‡] 94.39 (90.66)
	Meng et al. 17 [65]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-		8 folds	✗	7 classes[‡] 95.37 (95.51)
	Liu et al. 17 [78]	CNN	loss layer	11	-	IntraFace	✓	IN		8 folds	✗	7 classes[‡] 97.1 (96.1)
	Yang et al. 18 [105]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	7 classes[‡] 97.30 (96.57)
	Zhang et al. 18 [46]	CNN	$\mathcal{MN}$	-	-	✓	✓	-		10 folds	✗	6 classes: 98.9
JAFFE	Liu et al. 14 [13]	DBN	$\mathcal{CN}$	6	2m	✓	-	-	213 images	LOSO	AdaBoost	7 classes[‡] 91.8
	Hamester et al. 15 [106]	CNN, CAE[2]	$\mathcal{NE}$	3	-	-	-	IN			✗	7 classes[‡] (95.8)
MMI	Liu et al. 13 [101]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-	the middle three frames and the first frame	10 folds	SVM	7 classes[‡] 74.76 (71.73)
	Liu et al. 15 [102]	CNN, RBM	$\mathcal{CN}$	5	-	V&J	-	-		10 folds	SVM	7 classes[‡] 75.85
	Mollahosseini et al. 16 [14]	CNN (Inception)		11	7.3m	IntraFace	✓	-	images from each sequence	5 folds	✗	6 classes: 77.9
	Liu et al. 17 [78]	CNN	loss layer	11	-	IntraFace	✓	IN	the middle three frames	10 folds	✗	6 classes: 78.53 (73.50)
	Li et al. 17 [34]	CNN	loss layer	8	5.8m	IntraFace	✓	-		5 folds	SVM	6 classes: 78.46
	Yang et al. 18 [105]	GAN (cGAN)		-	-	MoT	✓	-		10 folds	✗	6 classes: 73.23 (72.67)
	Liu et al. 19 [107]	CNN	loss layer	-	-	IntraFace	✓	IN		10 folds	✗	6 classes: 81.13 (79.33)
TFD	Reed et al. 14 [108]	RBM	$\mathcal{MN}$	-	-	-	-	-	4,178 emotion labeled 3,874 identity labeled	5 official folds	SVM	Test: 85.43
	Devries et al. 14 [61]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN	4,178 labeled images		✗	Validation: 87.80 Test: 85.13 (48.29)
	Khorrami et al. 15 [103]	zero-bias CNN		4	7m	✓	✓	-			✗	Test: 88.6
	Ding et al. 17 [68]	CNN	fine-tune	8	11m	IntraFace	✓	-			✗	Test: 88.9 (87.7)
FER 2013	Tang 13 [93]	CNN	loss layer	4	12.0m	-	✓	IN	Training Set: 28,709 Validation Set: 3,589 Test Set: 3,589		✗	Test: 71.2
	Devries et al. 14 [61]	CNN	$\mathcal{MN}$	4	12.0m	MoT	✓	IN			✗	Validation+Test: 67.21
	Zhang et al. 15 [109]	CNN	$\mathcal{MN}$	6	21.3m	SDM	-	-			✗	Test: 75.10
	Guo et al. 16 [110]	CNN	loss layer	10	2.6m	SDM	✓	-			k-NN	Test: 71.33
	Kim et al. 16 [111]	CNN	$\mathcal{NE}$	5	2.4m	IntraFace	✓	IN			✗	Test: 73.73
	Pramerdorfer et al. 16 [112]	CNN	$\mathcal{NE}$	10/16/33	1.8/1.2/5.3 (m)	-	✓	IN			✗	Test:75.2
	Georgescu et al. 19 [113]	CNN	$\mathcal{NE}$	8/13/16	-	-	✓	-			SVM	Test:75.42
SFEW 2.0	levi et al. 15 [79]	CNN	$\mathcal{NE}$	VGG-S/VGG-M/ GoogleNet		MoT	✓	-	891 training, 431 validation, and 372 test		✗	Validation: 51.75 Test: 54.56
	Ng et al. 15 [67]	CNN	fine-tune	AlexNet		IntraFace	✓	-	921 training, ? validation, and 372 test		✗	Validation: 48.5 (39.63) Test: 55.6 (42.69)
	Li et al. 17 [34]	CNN	loss layer	8	5.8m	IntraFace	✓	-	921 training, 427 validation		SVM	Validation: 51.05
	Ding et al. 17 [68]	CNN	fine-tune	8	11m	IntraFace	✓	-	891 training, 425 validation		✗	Validation: 55.15 (46.6)
	Liu et al. 17 [78]	CNN	loss layer	11	-	IntraFace	✓	IN	958training, 436 validation, and 372 test		✗	Validation: 54.19 (47.97)
	Cai et al. 17 [104]	CNN	loss layer	6	-	DRMF	✓	IN			✗	Validation: 52.52 (43.41) Test: 59.41 (48.29)
	Meng et al. 17 [65]	CNN	$\mathcal{MN}$	6	-	DRMF	✓	-			✗	Validation: 50.98 (42.57) Test: 54.30 (44.77)
	Kim et al. 15 [54]	CNN	$\mathcal{NE}$	5	-	multiple	✓	IN			✗	Validation: 53.9 Test: 61.6
	Yu et al. 15 [53]	CNN	$\mathcal{NE}$	8	6.2m	multiple	✓	IN			✗	Validation: 55.96 (47.31) Test: 61.29 (51.27)

[1] The value in parentheses is the mean accuracy, which is calculated with the confusion matrix given by the authors.
† 7 Classes: anger, contempt, disgust, fear, happiness, sadness, and surprise.
‡ 7 Classes: anger, disgust, fear, happiness, neutral, sadness, and surprise.

expression-discriminative ability of the learned features. To eliminate this effect, a two-stage training algorithm FaceNet2ExpNet [68] was proposed. The fine-tuned face net serves as a suitable initialization for the expression net and is used to supervise the training of the convolutional layers only. The fully connected layers are trained from scratch with expression information to regularize the training of the target FER net.

4.2 Diverse network input

Traditional practices commonly use the whole aligned face of RGB images as the input of the network to learn features for FER. However, these raw data lack important information, such as homogeneous or regular textures and invariance in terms of image scaling, rotation, occlusion and illumination, which may represent confounding factors for FER. Some methods have employed diverse handcrafted features and their extensions as the network input to strengthen the network's robustness to common distractions and to force the network to focus more on facial areas with expressive information.

Low-level representations encode features from small regions in the given RGB image and then cluster and pool these features with local histograms, which are robust to illumination variations and small registration errors. A novel mapped LBP feature [79] was proposed for illumination-invariant FER. Scale-invariant feature transform (SIFT) [118]) features that are robust against image scaling and rotation are employed [119] for multiview FER tasks. Combining different descriptors in outline, texture, angle, and color as the input data can also help enhance the deep network performance [58], [120].

In addition, *part-based representations* extract features according to the target task, which remove noncritical parts from the whole image and exploit key parts that are sensitive to the task. [121] indicated that three regions of interest (ROIs), i.e., eyebrows, eyes and mouth, are strongly related to facial expression changes. Other studies [122], [123], [124], [125], [126], [127] proposed automatically learning the key parts (salient features) for facial expressions using attention mechanisms.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

4.3 Deep FER networks for static images

A large volume of the existing studies conducted expression recognition tasks based on static images without considering temporal information due to the convenience of data processing and the availability of related training and test data. For each of the most frequently evaluated datasets, Table 3 shows the current state-of-the-art methods in the field that are explicitly conducted in a person-independent protocol.

4.3.1 Auxiliary block

Based on the foundation architecture of the CNN, several studies have proposed the addition of well-designed auxiliary blocks or layers to enhance the expression-related representation capability of learned features.

A novel CNN architecture, HoloNet [88], was designed for FER, where CReLU [128] was combined with the residual structure to increase the network depth without efficiency reduction and an inception-residual block was uniquely designed for FER to learn multiscale and expression-discriminative features. Another CNN model, supervised scoring ensemble (SSE) [89], was introduced to enhance the supervision degree for FER, where three kinds of supervised blocks were embedded in the early hidden layers of the mainstream CNN for shallow, intermediate and deep supervision (see Fig. 4). Interestingly, Zeng et al. [129] noted that inconsistent annotations among different FER databases are inevitable, which would damage the performance when the training set is enlarged by merging multiple datasets. To address this problem, the authors proposed an Inconsistent Pseudo Annotations to Latent Truth (IPA2LT) framework. In IPA2LT, an end-to-end trainable LTNet is designed to discover the latent truths from the human annotations and the machine annotations trained from different datasets by maximizing the log-likelihood of these inconsistent annotations.

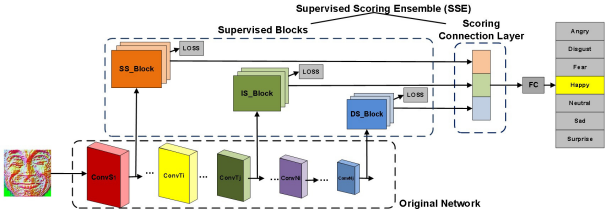
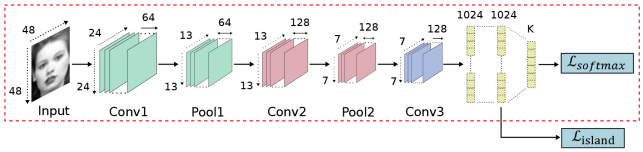


Fig. 4. Three different supervised blocks in [89]. SS_Block for shallow-layer supervision, IS_Block for intermediate-layer supervision, and DS_Block for deep-layer supervision. These blocks were designed according to the layerwise feature description ability of the original network.

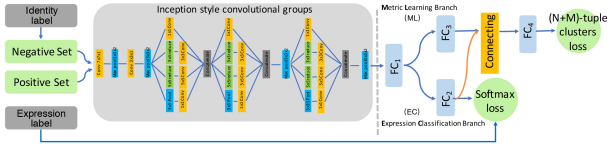
4.3.2 Loss layer

The traditional softmax loss layer in CNNs simply forces features of different classes to remain apart, but FER in realistic conditions suffers from not only high interclass similarity but also high intraclass variation. Therefore, several works have proposed novel loss layers to mitigate this problem.

Inspired by the center loss [130], which penalizes the distance between deep features and their corresponding class centers, two variations were proposed to assist the supervision of the softmax loss for more discriminative features. In [104], island loss was formalized to increase the pairwise distances between different class centers (see Fig. 5(a)). Specifically, the island loss calculated at the feature extraction layer and the softmax loss calculated at the decision layer are combined to supervise the CNN training. In



(a) Island loss layer in [104].



(b) (N+M)-tuple clusters loss layer in [78].

Fig. 5. Representative loss layers that are specifically designed for deep facial expression recognition.

[34], locality-preserving loss (LP loss) was formalized to pull the locally neighboring features of the same class together so that the intraclass local clusters can be closer for each expression. Jointly training this loss with the softmax loss, the discriminative power of the learned features can be highly enhanced.

Based on the triplet loss [131], which requires one positive example to be closer to the anchor than one negative example with a fixed gap, two variations were proposed to replace or assist the supervision of the softmax loss. In [110], exponential triplet-based loss was formalized to give difficult samples more weight when updating the network. In [78], (N+M)-tuples cluster loss was formalized to alleviate the difficulty of anchor selection and threshold validation in the triplet loss for identity-invariant FER (see Fig. 5(b)). During training, identity-aware hard-negative mining and online positive mining schemes were used to decrease the inter-identity variation in the same expression.

4.3.3 Network ensemble

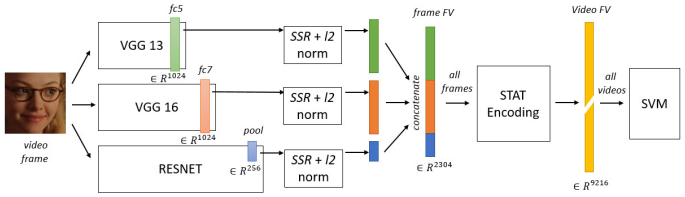
Previous research has suggested that assemblies of multiple networks can outperform an individual network [132]. Two key factors should be considered when implementing network ensembles: (1) sufficient diversity of the networks to ensure complementarity, and (2) an appropriate ensemble method that can effectively aggregate the committee networks.

In terms of the first factor, different kinds of training data and various network parameters or architectures are considered to generate diverse committees. Several preprocessing methods [111], such as deformation and normalization, and the methods described in Section 4.2 can generate different data to train diverse networks. By changing the size of filters, the number of neurons and the number of layers in the networks, and applying multiple random seeds for weight initialization, the diversity of the networks can also be enhanced [54], [133]. In addition, different architectures of networks can be used to enhance diversity. For example, a supervised CNN and an unsupervised convolutional autoencoder were combined for the network ensemble in [106].

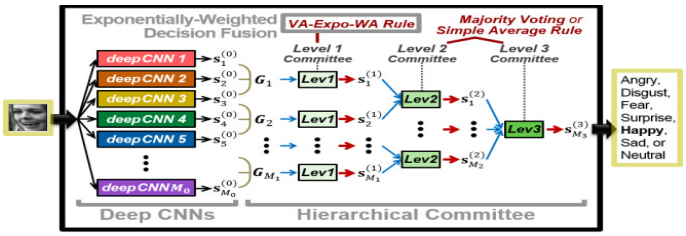
For the second factor, each member of the committee networks can be assembled at two different levels: the feature level and the decision level. For *feature-level* ensembles, the most commonly adopted strategy is to concatenate features learned from different networks [85], [113], [134]. For example, [85] concatenated features learned from different networks to obtain a single feature vector to describe the input image (see Fig. 6(a)). In addition, a feature loss [135] was proposed to embed the information

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing



(a) Feature-level ensemble in [85].



(b) Decision-level ensemble in [54].

Fig. 6. Representative network ensemble systems at the feature level and decision level. (a) Three different features (*fc5* of VGG13 + *fc7* of VGG16 + *pool* of Resnet) after normalization are concatenated to obtain a single feature vector (FV) that describes the input frame. (b) A 3-level hierarchical committee architecture with hybrid decision-level fusions was proposed to obtain sufficient decision diversity.

of handcrafted features into the training process and provide complementary information for the deeply learned features. For *decision-level* ensembles, three widely used rules are applied: majority voting, simple average and weighted average. Because the weighted average rule considers the importance and confidence of each individual, many weighted average methods have been proposed to find an optimal set of weights for network ensembles (e.g., random search [60], log-likelihood loss and hinge loss [53], exponentially weighted average in Fig. 6(b) [54] and trainable CNN [133]).

4.3.4 Multitask networks

Many existing networks for FER focus on a single task and learn features that are sensitive to expressions without considering interactions among other latent factors. However, in the real world, FER is intertwined with various factors, such as head pose, illumination, and subject identity (facial morphology). To solve this problem, multitask learning is introduced to transfer knowledge from other relevant tasks and to disentangle nuisance factors.

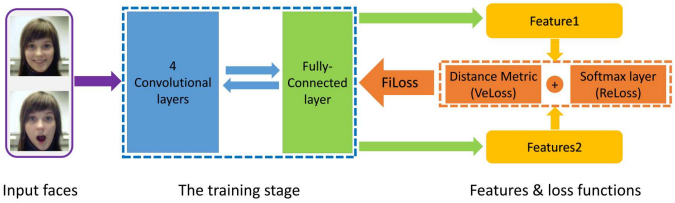


Fig. 7. Representative multitask network for FER. In the proposed MSCNN [73], a pair of images is sent into the MSCNN during training. The expression recognition task with cross-entropy loss, which learns features with large between-expression variation, and the face verification task with contrastive loss, which reduces the variation in within-expression features, are combined to train the MSCNN.

Reed et al. [108] constructed a higher-order Boltzmann machine (disBM) to learn manifold coordinates for the relevant factors of expressions and proposed training disentangling strategies so that the expression-related hidden units are invariant to face morphology. Other works [61], [136] suggested that simultaneously conducting FER with other tasks, such as facial landmark localization and facial AU [137] detection, can jointly improve FER performance. In addition, several works [65], [73] employed multitask learning for *identity-invariant FER*. In [65], an identity-aware CNN (IACNN) with two identical subCNNs was proposed. One stream used expression-sensitive loss for expression-discriminative features, and the other stream used the identity-sensitive loss to learn identity-related features for identity-invariant FER. In [73], a multisignal CNN (MSCNN), which was trained under the supervision of both FER and face verification tasks, was proposed to force the model to focus on expression information (see Fig. 7). Furthermore, an all-in-one CNN model [138] was proposed to simultaneously solve a diverse set of face analysis tasks, including smile detection. Similarly, SmileNet [139] was proposed to learn both face detection and smile recognition, which does not require a prenormalization step including face detection and registration.

The conventional supervised multitask learning mentioned above requires training samples labeled for all tasks. To address this, [46] proposed a novel attribute propagation method that can leverage the inherent correspondences between facial expressions and other heterogeneous attributes despite the disparate distributions of different datasets.

4.3.5 Cascaded Networks

In a cascaded network, various modules for different tasks are combined sequentially to construct a deeper network, where the outputs of the former modules are utilized by the latter modules. Related studies have proposed combinations of different structures to learn a hierarchy of features through which factors of variation that are unrelated to expressions can be gradually filtered out.

Most commonly, different networks or learning methods are combined sequentially and individually, and each of them contributes differently and hierarchically. In [140], DBNs were trained to first detect face expression-related areas. Then, these parsed face components were classified by a stacked autoencoder. In [141], a multiscale contractive convolutional network (CCNET) was proposed to obtain local-translation-invariant (LTI) representations. Then, a contractive autoencoder was designed to hierarchically separate the emotion-related factors from subject identity and pose. In [101], [102], overcomplete representations were first learned using the CNN architecture, and then a multilayer RBM was exploited to learn higher-level features for FER (see Fig. 8).

Instead of simply concatenating the outputs of different networks, Liu et al. [13] presented a boosted DBN (BDBN) that iteratively performs feature representation, feature selection and classifier construction in a unified loopy state. Compared with the concatenation without feedback, this loopy framework propagates the classification error backward to initiate the feature selection process alternately until convergence. Thus, the discriminative ability for FER can be substantially improved during this iteration.

4.3.6 Generative adversarial networks (GANs)

Recently, GAN-based methods have been successfully used in image synthesis to generate impressively realistic faces, numbers, and a variety of other image types, which are beneficial to training data augmentation and the corresponding recognition tasks.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

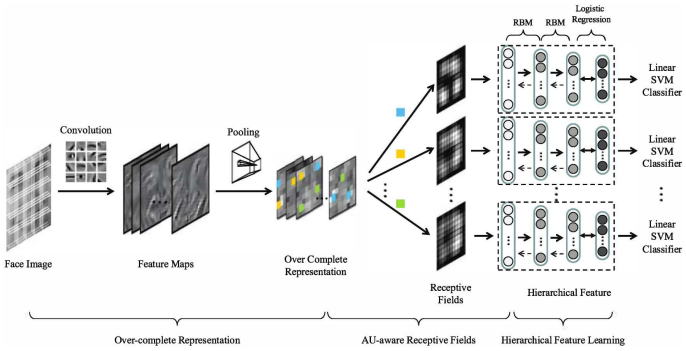


Fig. 8. Representative cascaded network for FER. The proposed AU-aware deep network (AUDN) [101] is composed of three sequential models: the first model trains a 2-layer CNN to obtain an overcomplete representation encoding all expression-discriminative variations in all possible locations; the second model contains an AU-aware receptive field layer that searches subsets of the overcomplete representation; and the last model is a multilayer RBM that learns hierarchical features.

Several works have proposed novel GAN-based models for pose-invariant FER and identity-invariant FER.

For pose-invariant FER, Lai et al. [142] proposed a GAN-based face frontalization framework, where the generator frontalizes input face images while preserving the identity and expression characteristics and the discriminator distinguishes the real images from the generated frontal face images. Zhang et al. [143] proposed a GAN-based model that can generate images with different expressions under arbitrary poses for multiview FER. For identity-invariant FER, Yang et al. [144] proposed an identity-adaptive generation (IA-gen) model with two parts. The upper part generates images of the same subject with different expressions using cGANs. Then, the lower part conducts FER for each single-identity subspace without involving other individuals; thus, identity variations can be well alleviated. Chen et al. [145] proposed a privacy-preserving representation-learning variational GAN (PPRL-VGAN) that combines VAE and GAN to learn an identity-invariant representation that is explicitly disentangled from the identity information and generative for expression-preserving face image synthesis. Yang et al. [105] proposed a de-expression residue learning (DeRL) procedure for exploring the expressive information, which is filtered out during the de-expression process but still embedded in the generator. Then, the model extracts this information from the generator directly to mitigate the influence of subject variations and improve the FER performance.

4.3.7 Discussion

The existing well-constructed deep FER systems focus on two key issues: the lack of plentiful diverse training data and expression-unrelated variations, such as illumination, head pose and identity. Table 4 shows the relative advantages and disadvantages of these different types of methods with respect to two open issues (data size requirement and expression-unrelated variations) and other focuses (computation efficiency, performance and difficulty of network training).

Instead of the popular network architecture, various *auxiliary blocks* and *loss layers* are specifically designed for FER to enhance the supervision degree of the network and to learn more powerful features with discriminate interclass separability and intraclass compactness. However, additional blocks may influence the computational efficiency of the whole network. It takes time to learn extra hyperparameters for the new loss and find a proper trade-off between different loss layers.

TABLE 4
Comparison of different types of methods for static images in terms of data size requirement, variations* (head pose, illumination, occlusion and other environment factors), identity bias, computational efficiency, accuracy, and difficulty on network training.

Network type	data	variations*	identity bias	efficiency	accuracy	difficulty
Auxiliary block	varies	good	varies	varies	good	varies
Loss layer	fair	good	varies	varies	good	varies
Network ensemble	low	good	fair	low	good	medium
Multitask network	high	varies	good	fair	varies	hard
Cascaded network	fair	good	fair	fair	fair	medium
GAN	fair	good	good	fair	good	hard

Training a deep and wide network with a large number of hidden layers and flexible filters is an effective method for learning deep high-level features that are discriminative for the target task. However, this process is vulnerable to the size of the training data and can underperform if insufficient training data are available for learning the new parameters. Integrating multiple relatively small networks in parallel or in series is a natural research direction for overcoming this problem. *Network ensemble* integrates diverse networks at the feature or decision level to combine their advantages and is usually applied in emotion competitions to help improve the performance. However, designing different kinds of networks to compensate for each other obviously increases the computational cost and the storage requirement. Moreover, the weight of each subnetwork is usually learned according to the performance of the original training data, leading to overfitting on newly unseen testing data. *Multitask networks* jointly train multiple networks considering interactions between the target FER task and other secondary tasks, such as facial landmark localization, facial AU recognition and face verification; thus, the expression-unrelated factors, including identity bias, can be well separated. The downside to this method is that it requires labeled data from all tasks, and the training becomes increasingly cumbersome as more tasks are involved. Alternatively, *cascaded networks* sequentially train multiple networks in a hierarchical approach, in which case the discriminative ability of the learned features is continuously strengthened. In general, this method can alleviate the overfitting problem and progressively separate factors that are irrelevant to facial expression. A deficiency worth considering is that the subnetworks in most existing cascaded systems are trained individually without feedback, and the end-to-end training strategy is preferable for enhancing training effectiveness and performance [13].

Ideally, deep networks, especially CNNs, are good at solving head-pose variations, yet most current FER networks have not explicitly addressed these variations, and they have not been tested in realistic conditions. *Generative adversarial networks (GANs)* can be exploited to solve this issue by frontalizing face images while preserving expression characteristics [142] or synthesizing arbitrary poses to help train the pose-invariant network [143]. Another advantage of GANs is that identity variations can be explicitly disentangled by generating the corresponding neutral face image [105] or synthesizing different expressions while preserving identity information for identity-invariant FER [144].

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

Moreover, GANs can help augment the training data on both size and diversity. The main drawback of GANs is the training instability and the trade-off between visual quality and image diversity.

4.4 Deep FER networks for dynamic image sequences

Although most of the previous models focus on static images, FER can benefit from the temporal correlations of consecutive frames in a sequence. We first introduce the existing frame aggregation techniques that strategically combine deep features learned from static-based FER networks. Then, considering that in a videostream, people usually display the same expression with different intensities, we further review methods that use images in different expression intensity states for intensity-invariant FER. Finally, we introduce deep FER networks that consider spatiotemporal motion patterns in video frames and learned features derived from the temporal structure. For each of the most frequently evaluated datasets, Table 5 shows the current state-of-the-art methods conducted in the person-independent protocol.

4.4.1 Frame aggregation

Because the frames in a given video clip may vary in expression intensity, directly measuring per-frame error does not yield satisfactory performance. Various methods have been proposed to aggregate the network output for frames in each sequence to improve performance. We divide these methods into two groups: decision-level frame aggregation and feature-level frame aggregation.

For decision-level frame aggregation, n-class probability vectors of each frame in a sequence are integrated. The most convenient method is to directly concatenate the output of these frames. However, the number of frames in each sequence may be different. Two aggregation approaches have been considered to generate a fixed-length feature vector for each sequence [60], [146]: frame averaging and frame expansion. An alternative approach that does not require a fixed number of frames is applying statistical coding. The average, max, average of square, average of maximum suppression vectors and so on can be used to summarize the per-frame probabilities in each sequence.

For feature-level frame aggregation, the learned features of frames in the sequence are aggregated. Many statistical-based encoding modules can be applied in this scheme. A simple and effective method is to concatenate the mean, variance, minimum, and maximum of the features over all frames [85]. Alternatively, matrix-based models such as eigenvectors, covariance matrices and multi-dimensional Gaussian distributions can also be employed for aggregation [147], [148]. In addition, multi-instance learning has been explored for video-level representation [149], where the cluster centers are computed from auxiliary data and then a bag-of-words representation is obtained for each bag of video frames.

4.4.2 Expression intensity-invariant network

Most methods (introduced in Section 4.3) focus on recognizing the peak high-intensity expression and ignore the subtle lower-intensity expressions. In this section, we introduce expression intensity-invariant networks that take training samples with different intensities as input to exploit the intrinsic correlations among expressions from a sequence that vary in different intensities.

In an expression intensity-invariant network, image frames with intensity labels are used for training. During the test, data

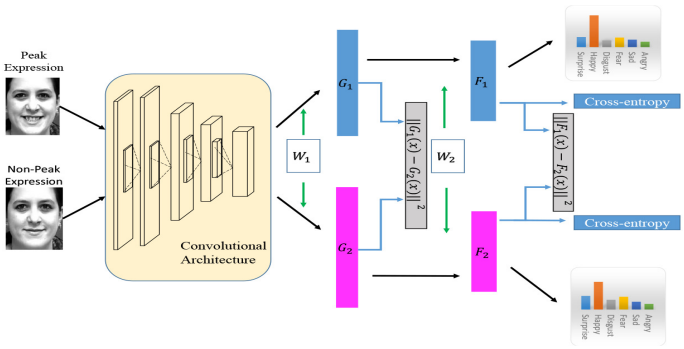


Fig. 9. The proposed PPDN in [17]. During training, PPDN is jointly optimized by the L2-norm loss and the cross-entropy loss of two expression images. During testing, the PPDN takes one still image as input for probability prediction.

that vary in expression intensity are used to verify the intensity-invariant ability of the network. Zhao et al. [17] proposed a peak-piloted deep network (PPDN) that takes a pair of peak and nonpeak images with the same expression and subject identity as input and utilizes the L2-norm loss to minimize the distance between both images. During backpropagation, peak gradient suppression (PGS) was proposed to drive the learned feature of the nonpeak expression towards that of the peak expression while avoiding the inverse. Thus, the network discriminant ability on lower-intensity expressions can be improved (see Figure 9). Based on PPDN, Yu et al. [75] proposed a deeper cascaded peak-piloted network (DCPN) that used a deeper and larger architecture to improve the discriminative ability of the learned representations and employed an integration training method called cascade fine-tuning to avoid overfitting. In [71], more intensity states (onset, onset to apex transition, apex, apex to offset transition and offset) were utilized, and five loss functions were adopted to regulate the network training by minimizing expression classification error, intraclass expression variation, intensity classification error and intra-intensity variation, and encoding intermediate intensity.

Considering that images with different expression intensities are not always available in the real world, several works have proposed automatically acquiring the intensity label or generating new images with targeted intensity. For example, in [159], the peak and neutral frames were automatically selected from the sequence in two stages: a clustering stage to divide all frames into the peak-like group and the neutral-like group using the k-means algorithm and a classification stage to detect peak and neutral frames using a semisupervised SVM. In [150], a deep generative-contrastive model was presented with two steps: a generator to generate the reference (less-expressive) face for each sample via a convolutional encoder-decoder and a contrastive network to jointly filter out information that is irrelevant to expressions through a contrastive metric loss and a supervised reconstruction loss.

4.4.3 Deep spatiotemporal FER network

Although frame aggregation can integrate frames in the video sequence, the crucial temporal dependency is not explicitly exploited. In contrast, the spatiotemporal FER network takes a series of frames in a temporal window as an independent input without prior knowledge of the expression intensity and utilizes both textural and temporal information to encode more subtle expressions.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

TABLE 5
Performances of representative methods for dynamic-based deep facial expression recognition on the most widely evaluated datasets. Network size = depth & number of parameters; Preprocessing = Face detection & Data augmentation & Face normalization; IN = illumination normalization; $\mathcal{FA}$ = frame aggregation; $\mathcal{EIN}$ = expression intensity-invariant network; $\mathcal{FLT}$ = facial landmark trajectory; $\mathcal{CN}$ = cascaded network; $\mathcal{NE}$ = network ensemble; S = spatial network; T = temporal network; LOSO = leave-one-subject-out.

Datasets	Methods	Network type	Network size		Preprocessing			Training data selection in each sequence	Testing data selection in each sequence	Data group	Performance[1](%)
CK+	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 99.3
	Yu et al. 17 [75]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 99.6
	kim et al. 17 [150]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames	the same as the training data	10 folds	7 classes: 97.93
	Sun et al. 17 [151]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 97.28
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 92.35
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 91.44
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames		10 folds	7 classes: 97.25 (95.22)
	Kumawat et al. 19 [152]	C3D	-	1.6m	-	✓	-	fixed length 11		10 folds	7 classes: 97.38 (96.65)
	kuo et al. 18 [87]	$\mathcal{FA}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	7 classes: 98.47
	Zhang et al. 17 [73]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	7 classes: 98.50 (97.78)
MMI	Kim et al. 17 [71]	$\mathcal{EIN}$, $\mathcal{CN}$	7	1.5m	Incremental	✓	-	5 intensities frames	the same as the training data	LOSO	6 classes: 78.61 (78.00)
	kim et al. 17 [150]	$\mathcal{EIN}$	14	-	✓	✓	-	all frames		10 folds	6 classes: 81.53
	Hasani et al. 17 [153]	$\mathcal{FLT}$, $\mathcal{CN}$	22	-	3000 fps	-	-	ten frames		5 folds	6 classes: 77.50 (74.50)
	Hasani et al. 17 [59]	$\mathcal{CN}$	29	-	AAM	-	-	static frames		5 folds	6 classes: 78.68
	Zhang et al. 17 [73]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the middle frame T: all frames		10 folds	6 classes: 81.18 (79.30)
	Wang et al. 19 [154]	$\mathcal{FLT}$	-	-	SDM	✓	-	fixed number of frames		10 folds	6 classes: 82.21
	Sun et al. 17 [151]	$\mathcal{NE}$	3 * GoogLeNetv2		✓	-	-	S: emotional T: neutral+emotional		10 folds	6 classes: 91.46
Oulu-CASIA	Zhao et al. 16 [17]	$\mathcal{EIN}$	22	6.8m	✓	-	-	from the 7th to the last[2]	the last frame	10 folds	6 classes: 84.59
	Yu et al. 17 [75]	$\mathcal{EIN}$	42	-	MTCNN	✓	-	from the 7th to the last[2]	the peak expression	10 folds	6 classes: 86.23
	Jung et al. 15 [16]	$\mathcal{FLT}$	2	177.6k	IntraFace	✓	-	fixed number of frames	the same as the training data	10 folds	6 classes: 74.17
	Jung et al. 15 [16]	C3D	4	-	IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 74.38
	Jung et al. 15 [16]	$\mathcal{NE}$	$\mathcal{FLT}$/C3D		IntraFace	✓	-	fixed number of frames		10 folds	6 classes: 81.46 (81.49)
	Kumawat et al. 19 [152]	C3D	-	1.6m	-	✓	-	fixed length 11		10 folds	6 classes: 82.41 (82.41)
	Zhang et al. 17 [73]	$\mathcal{NE}$	7/5	2k/1.6m	SDM/ Cascaded CNN	✓	-	S: the last frame T: all frames		10 folds	6 classes: 86.25 (86.25)
	kuo et al. 18 [87]	$\mathcal{NE}$	6	2.7m	IntraFace	✓	IN	fixed length 9		10 folds	6 classes: 91.67
AFEW* 6.0	Ding et al. 16 [148]	$\mathcal{FA}$	AlexNet		✓	-	-	Training: 773; Validation: 373; Test: 593			Validation: 44.47
	Yan et al. 16 [155]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	40 frames		3 folds	7 classes: 44.46
	Yan et al. 16 [155]	$\mathcal{FLT}$	4	-	[156]	-	-	30 frames		3 folds	7 classes: 37.37
	Fan et al. 16 [157]	$\mathcal{CN}$	VGG16-LSTM		✓	-	-	16 features for LSTM			Validation: 45.43 (38.96)
	Fan et al. [157]	C3D	10	-	✓	-	-	several windows of 16 consecutive frames			Validation: 39.69 (38.55)
	Yan et al. 16 [155]	fusion	/					Training: 773; Validation: 383; Test: 593			Test: 56.66 (40.81)
	Fan et al. 16 [157]	fusion	/					Training: 774; Validation: 383; Test: 593			Test: 59.02 (44.94)
AFEW* 7.0	Ouyang et al. 17 [76]	$\mathcal{CN}$	VGG-LSTM		MTCNN	✓	-	16 frames			Validation: 47.4
	Ouyang et al. 17 [76]	C3D	10	-	MTCNN	✓	-	16 frames			Validation: 35.2
	Vielzeuf et al. [158]	$\mathcal{CN}$	C3D-LSTM		✓	✓	-	detected face frames			Validation: 43.2
	Vielzeuf et al. [158]	$\mathcal{CN}$	VGG16-LSTM		✓	✓	-	several windows of 16 consecutive frames			Validation: 48.6
	Vielzeuf et al. [158]	fusion	/					Training: 773; Validation: 383; Test: 653			Test: 58.81 (43.23)

[1] The value in parentheses is the mean accuracy calculated from the confusion matrix given by the authors.
[2] A pair of images (peak and nonpeak expression) is chosen for training each time.
* We included the result of a single spatiotemporal network and the best result after fusion with both video and audio modalities.
† 7 Classes in CK+: anger, contempt, disgust, fear, happiness, sadness, and surprise.
‡ 7 Classes in AFEW: anger, disgust, fear, happiness, neutral, sadness, and surprise.

RNN and C3D: RNN can robustly derive information from sequences by exploiting the fact that feature vectors for successive data are connected semantically and are therefore interdependent. The improved version, LSTM, is flexible to handle varying-length sequential data with lower computation cost. Derived from RNN, an RNN that is composed of ReLUs and initialized with the identity matrix (IRNN) [160] was used to provide a simple mechanism to address the exploding and vanishing gradient problems [84]. Bidirectional RNNs (BRNNs) [161] were employed to learn the temporal relations in both the original and reversed directions [73], [155]. Recently, a nested LSTM [77] was proposed with two sub-LSTMs. Namely, T-LSTM models the temporal dynamics of the learned features, and C-LSTM integrates the outputs of all T-LSTM models to obtain the multilevel representations. [162] employed ConvLSTM with a 2D grid convolution to encode the spatial correlations and model spatiotemporal relationships for the input expression sequences.

Compared with RNN, CNN is more suitable for computer vision applications; hence, its derivative C3D [163], which uses 3D convolutional kernels with shared weights along the time axis instead of the traditional 2D kernels, has been widely used for dynamic-based FER (e.g., [76], [81], [157], [162], [164], [165]) to capture the spatiotemporal features. Based on C3D, many derived structures have been designed for FER. In [166], 3D CNN was incorporated with the DPM-inspired [167] deformable facial action constraints to simultaneously encode dynamic motion and discriminative part-based representations. In [16], a deep temporal appearance network (DTAN) was proposed that employed 3D filters without sharing weights along the timeline; hence, each filter can vary in importance over time. Likewise, a weighted C3D was proposed [158], where several windows of consecutive frames were extracted from each sequence and weighted based on their prediction scores. Instead of directly using C3D for classification, [168] employed C3D for spatiotemporal feature extraction and then cascaded it with DBN for prediction. In [169], C3D was also used as a feature extractor, followed by a NetVLAD layer [170] to aggregate the temporal information of the motion features by learning cluster centers.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

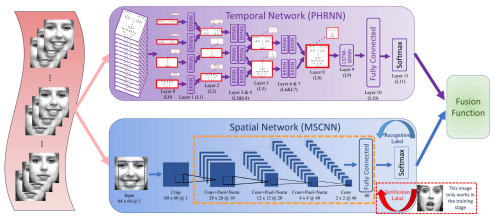


Fig. 10. The spatiotemporal network in [73]. The temporal network PHRNN for landmark trajectory and the spatial network MSCNN for identity-invariant features are trained separately. The predicted probabilities from the two networks are fused together for spatiotemporal FER.

Facial landmark trajectory: Related psychological studies have shown that expressions are invoked by dynamic motions of certain facial parts (e.g., eyes, nose and mouth) that contain the most descriptive information for representing expressions. To obtain more accurate facial actions for FER, facial landmark trajectory models have been proposed to capture the dynamic variations of facial components from consecutive frames.

To extract landmark trajectory representation, the most direct way is to concatenate coordinates of facial landmark points from frames over time with normalization to generate a one-dimensional trajectory signal for each sequence [16] or to form an image-like map as the input of CNN [155]. In addition, the relative distance variation in each landmark in consecutive frames can also be used to capture the temporal information [171]. Furthermore, a part-based model that divides facial landmarks into several parts according to the facial physical structure and then separately feeds them into the networks hierarchically is proven to be efficient for both local low-level and global high-level feature encoding [73] (see "PHRNN" in Fig. 10). Instead of separately extracting the trajectory features and then inputting them into the networks, Hasani et al. [153] incorporated the trajectory features by replacing the shortcut in the residual unit of the original 3D Inception-ResNet with elementwise multiplication of facial landmarks and the input tensor of the residual unit. Thus, the landmark-based network can be trained end-to-end.

Cascaded networks: By combining the powerful perceptual vision representations learned from CNNs with the strength of LSTM for variable-length inputs and outputs, Donahue et al. [172] proposed both spatially and temporally deep model that cascades the outputs of CNNs with LSTMs for various vision tasks involving time-varying inputs and outputs. Similar to this hybrid network, many cascaded networks have been proposed for FER (e.g., [71], [157], [162], [173]).

In addition to concatenating LSTM with the fully connected layer of CNN, a hypercolumn-based system [174] extracted the last convolutional layer features as the input of the LSTM for longer range dependencies without losing global coherence. Most CNN-LSTM methods train models that require the prediction to wait until the full sequence is available and may cause delays at test time. Hence, [175] proposed an on-the-fly prediction network that can learn spatiotemporal features with partial expression sequences and achieve higher recognition rates.

In addition to CNN, other network frameworks can also be used to learn spatial features, such as convolutional sparse autoencoder [176], 3D Inception-ResNet (3DIR) [153] and weighted C3D [158]. Likewise, in addition to LSTM, the conditional random fields model [59] was employed to distinguish the temporal relations of the input sequences.

TABLE 6
Comparison of different types of methods for dynamic image sequences in terms of data size requirement, representability of spatial and temporal information, requirement on frame length, performance, and computational efficiency. $\mathcal{FLT}$ = Facial Landmark Trajectory; $\mathcal{CN}$ = Cascaded Network; $\mathcal{NE}$ = Network Ensemble.

Network type		data	spatial	temporal	frame length	accuracy	efficiency
Frame aggregation		low	good	no	depends	fair	high
Expression intensity		fair	good	low	fixed	fair	varies
Spatio-temporal network	RNN	low	low	good	variable	low	fair
	C3D	high	good	fair	fixed	low	fair
	$\mathcal{FLT}$	fair	fair	fair	fixed	low	high
	$\mathcal{CN}$	high	good	good	variable	good	fair
	$\mathcal{NE}$	low	good	good	fixed	good	low

Network ensemble: A two-stream CNN for action recognition in videos, which trained one stream of the CNN on the multiframe dense optical flow for temporal information and the other stream of the CNN on still images for appearance features and then fused the outputs of two streams, was introduced by Simonyan et al. [177]. Inspired by this architecture, several network ensemble models have been proposed for FER.

Sun et al. [151] proposed a multichannel network that extracted the spatial information from emotion-expressing faces and temporal information (optical flow) from the changes between emotional and neutral faces and investigated three feature fusion strategies: score average fusion, SVM-based fusion and neural-network-based fusion. Zhang et al. [73] fused the temporal network PHRNN (discussed in **"Landmark trajectory"**) and the spatial network MSCNN (discussed in section 4.3.4) to extract the partial-whole, geometry-appearance, and static-dynamic information for FER (see Fig. 10). Instead of directly fusing the network outputs with precalculated weights that may cause overfitting problems in the testing phase, Jung et al. [16] proposed a joint fine-tuning method that jointly trained the DTAN (discussed in the **"RNN and C3D"**), the DTGN (discussed in the **"Landmark trajectory"**) and the integrated network, which outperformed the weighted sum strategy.

4.4.4 Discussion

In the real world, people display facial expressions in a dynamic process, e.g., from subtle to obvious, and it has become a trend to conduct FER on sequence/video data. Table 6 summarizes the relative merits of different types of methods on dynamic data in regards to the capability of representing spatial and temporal information, the requirement on training data size and frame length (variable or fixed), the computational efficiency and the performance.

Frame aggregation is employed to combine the learned feature or prediction probability of each frame for a sequence-level result. The output of each frame can be simply concatenated (fixed-length frames are required in each sequence) or statistically aggregated to obtain video-level representation (variable-length frames processible). This method is computationally simple and can achieve moderate performance if the temporal variations of the target dataset are not complicated.

According to the fact that the expression intensity in a video sequence varies over time, the *expression intensity-invariant network* considers images with nonpeak expressions and further

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

exploits the dynamic correlations between peak and nonpeak expressions. Commonly, image frames with specific intensity states are needed for intensity-invariant FER.

Despite the advantages of these methods, *frame aggregation* handles frames without consideration of temporal information and subtle appearance changes, and *expression intensity-invariant networks* require prior knowledge of expression intensity, which is unavailable in real-world scenarios. By contrast, *Deep spatiotemporal networks* are designed to encode temporal dependencies in consecutive frames and have been shown to benefit from learning spatial features in conjunction with temporal features. *RNN and its variations (e.g., LSTM, IRNN and BRNN) and C3D* are foundational networks for learning spatiotemporal features. However, the performance of these networks is barely satisfactory. RNN is incapable of capturing powerful convolutional features. 3D filters in C3D are applied over very short video clips ignoring long-range dynamics. Additionally, training such a large network is computationally a problem, especially for dynamic FER where video data are insufficient. Alternatively, *facial landmark trajectory* methods extract shape features based on the physical structures of facial morphological variations to capture dynamic facial component activities and then apply deep networks for classification. This method is computationally simple and can eliminate illumination variations. However, it is sensitive to registration errors and requires accurate facial landmark detection, which is difficult to access in unconstrained conditions. Consequently, this method performs less well and is more suitable to complement appearance representations. *Network ensemble* is utilized to train multiple networks for both spatial and temporal information and then to fuse the network outputs in the final stage. Optic flow and facial landmark trajectory can be used as temporal representations to collaborate spatial representations. One of the drawbacks of this framework is the precomputing and storage consumption of optical flow or landmark trajectory vectors. Most related studies randomly selected fixed-length video frames as input, leading to the loss of useful temporal information. *Cascaded networks* were proposed to first extract discriminative representations for facial expression images and then input these features to sequential networks to reinforce the temporal information encoding. However, this model introduces additional parameters to capture sequence information, and the feature learning network (e.g., CNN) and the temporal information encoding network (e.g., LSTM) in current works are not trained jointly, which may lead to suboptimal parameter settings. Training in an end-to-end fashion is still a long road.

Compared with deep networks on static data, Table 3 and Table 5 demonstrate the powerful capability and popularity trend of deep spatiotemporal networks. For instance, comparison results on widely evaluated benchmarks (e.g., CK+ and MMI) illustrate that training networks based on sequence data and analyzing temporal dependency between frames can further improve the performance. Additionally, in the EmotiW challenge 2015, only one system employed deep spatio-networks for FER, whereas 5 of 7 reviewed systems in the EmotiW challenge 2017 relied on such networks.

5 Additional Open Issues

In addition to the most popular basic expression classification task reviewed above, we further introduce a few related issues that depend on deep neural networks and prototypical expression-related knowledge.

5.1 Occlusion and nonfrontal head pose

Occlusion and nonfrontal head pose, which may change the visual appearance of the original facial expression, are two major obstacles for automatic FER, especially in real-world scenarios.

For *facial occlusion*, Ranzato et al. [178] proposed a deep generative model that used DBNs to model pixel-level features. Cheng et al. [179] employed multilayer RBMs with a pretraining and fine-tuning process to compress features from the occluded facial parts. Xu et al. [180] concatenated high-level learned features transferred from two CNNs with the same structure but pretrained on different datasets with additive occluded samples. Recently, Li et al. [126] proposed a CNN with an attention mechanism (ACNN) that can perceive the occlusion regions of the face and focus on the most discriminative unoccluded regions.

For *multiview FER*, Zhang et al. [119] introduced a projection layer into the CNN that learned discriminative features by weighting different facial landmarks within 2D SIFT feature matrices without requiring facial pose estimation. Liu et al. [181] proposed a multichannel pose-aware CNN (MPCNN) that contains three cascaded parts to predict expression labels by minimizing the conditional joint loss of pose and expression recognition. In addition, generative adversarial network (GAN) technology has been employed in [142], [143] to generate facial images with different expressions under arbitrary poses for multiview FER.

5.2 FER on 3D static and dynamic data

Despite significant advances achieved in 2D FER, it fails to solve two main problems: illumination changes and pose variations [28]. 3D FER that uses 3D face shape models with depth information can capture subtle facial deformations, which are naturally robust to pose and lighting variations.

Depth images and videos [182], [183] record the intensity of facial pixels based on distance from a depth camera, which contains critical information of facial geometric relations. To emphasize the dynamic deformation patterns of facial expression motions, [184] explored the 4D FER (3D FER using dynamic data) using a dynamic geometrical image network. Moreover, [185] proposed estimating 3D expression coefficients from image intensities using CNN without requiring facial landmark detection. Thus, the model is highly robust to extreme appearance variations, including out-of-plane head rotations, scale changes, and occlusions. To further enhance the robustness to pose variations, [186] proposed a fast and light manifold CNN that enhances geometry representation and highlights the shape characteristics of expressions.

Recently, an increasing number of works have tended to combine 2D and 3D data to improve performance. Oyedotun et al. [187] employed CNN to jointly learn facial expression features from both RGB and depth map latent modalities. Li et al. [188] proposed a deep fusion CNN to explore multimodal 2D+3D FER. Specifically, six types of 2D facial attribute maps were first extracted from 3D face scans and were then jointly fed into the feature extraction and feature fusion subnets to learn the optimal combination weights of 2D and 3D facial representations. To improve this work, [189] proposed extracting deep features from different facial parts extracted from texture and depth images and then fusing these features together to interconnect them with feedback.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

5.3 Facial expression synthesis

Realistic *facial expression synthesis*, which can generate various facial expressions for interactive interfaces, is a hot topic. Susskind et al. [190] demonstrated that DBN has the capacity to capture the large range of variation in expressive appearance and can be trained on large but sparsely labeled datasets. In light of this work, [178], [191], [192] employed DBN with unsupervised learning to construct facial expression synthesis systems. Kaneko et al. [115] proposed a multitask deep network with state recognition and key-point localization to adaptively generate visual feedback to improve FER. With the recent success of deep generative models, such as variational autoencoders (VAEs), adversarial autoencoders (AAEs), and generative adversarial networks (GANs), a series of facial expression synthesis systems have been developed based on these models (e.g., [193], [194], [195], [196], [197] and [198]). Facial expression synthesis can also be applied to data augmentation without manually collecting and labeling large datasets. Masi et al. [199] employed CNN to synthesize new face images by increasing face-specific appearance variation, such as expressions within the 3D textured face model.

5.4 Visualization techniques

In addition to utilizing CNN for FER, several works (e.g., [103], [200], [201]) employed *visualization* techniques [202] on the learned CNN features to qualitatively analyze how the CNN contributes to the appearance-based learning process of FER and to qualitatively decipher which portions of the face yield the most discriminative information. The deconvolutional results all indicated that the activations of some particular filters on the learned features have strong correlations with the face regions that correspond to facial AUs.

5.5 Other novel problems

We further discuss several novel issues that have been approached on the basis of the prototypical expression categories and need wider exploration. Dominant and complementary emotions have been investigated in the FG2017 challenge [203] to recognize more detailed emotions than basic emotions and see how different dominant emotions influence the recognition of complementary emotions. Real versus fake expressed emotion recognition has been approached in the ChaLearn Looking at People Challenge [204] to determine whether an emotion is fake or not. Deep learning techniques have been thoroughly applied by the participants of these two challenges (e.g., [205], [206], [207]). Recently, the issue of facial expression similarity that better mimics human visual preferences has been explored in [208] for developing various applications, such as expression retrieval and emotion recognition.

6 Challenges and Future Directions

6.1 Facial expression datasets

As the FER literature shifts its main focus to the challenging in-the-wild environmental conditions, many researchers have committed to employing deep learning technologies to handle difficulties, such as illumination variation, occlusions, nonfrontal head poses, identity bias and the recognition of low-intensity expressions. Given that FER is a data-driven task and that training a sufficiently deep network to capture subtle expression-related deformations requires a large amount of training data, the major challenge that deep FER systems face is the lack of training data in terms of both quantity and quality.

Because people of different age ranges, cultures and genders display and interpret facial expression in different ways, an ideal facial expression dataset is expected to include abundant sample images with precise face attribute labels, not just expression but other attributes such as age, gender and ethnicity, which would facilitate related research on cross-age range, cross-gender and cross-cultural FER using deep learning techniques, such as multitask deep networks and transfer learning. In addition, although occlusion and multipose problems have received relatively wide interest in the field of deep face recognition, occlusion-robust and pose-invariant issues have received less attention in deep FER. One of the main reasons is the lack of a sufficient facial expression dataset with occlusion type and head-pose annotations.

On the other hand, accurately annotating a large volume of image data with the large variation and complexity of natural scenarios is an obvious impediment to the construction of expression datasets. A reasonable approach is to employ crowd-sourcing models [34], [45], [209] under the guidance of expert annotators. Additionally, a fully automatic labeling tool [43] refined by experts is an alternative to provide approximate but efficient annotations. In both cases, a subsequent reliable estimation or labeling learning process is necessary to filter out noisy annotations. In particular, few comparatively large-scale datasets that consider real-world scenarios and contain a wide range of facial expressions have recently become publicly available, i.e., EmotioNet [43], RAF-DB [34], [44] and AffectNet [45], and we anticipate that with advances in technology and the widespread of the Internet, more complementary facial expression datasets will be constructed to promote the development of deep FER.

6.2 Dataset bias and imbalanced distribution

Data bias and inconsistent annotations are very common among different facial expression datasets due to different collecting conditions and the subjectiveness of annotation. Recent studies commonly evaluate their algorithms within a specific dataset and can achieve satisfactory performance [210]. However, algorithms evaluated via within-database protocols lack generalizability on unseen test data, and the performance in cross-dataset settings is greatly deteriorated due to the existing discrepancies. Furthermore, because of the inconsistent expression annotations, FER performance cannot keep improving when enlarging the training data by directly merging multiple datasets [129]. Cross-database performance is an important evaluation criterion of the generalizability and practicability of a FER system. Deep domain adaptation and knowledge distillation are promising trends to address this bias [211], [212].

Another common issue is imbalanced class distribution in facial expressions, which is a result of the practicality of sample acquirement. For example, collecting and annotating a happy face is simple; however, identifying the signals of disgust, fear and other less common expressions can be very laborious. As shown in Table 3 and Table 5, the performance assessed in terms of mean accuracy, which assigns equal weights to all classes, decreases when compared with the accuracy criterion, and this decline is especially evident in real-world datasets (e.g., SFEW 2.0 and AFEW). One solution is to resample and balance the class distribution based on the number of samples for each class during the preprocessing stage using data augmentation and synthesis

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

[213]. Another alternative is to develop a cost-sensitive loss layer for reweighting during network work training.

6.3 Incorporating other affective models

Another major issue that requires consideration is that while FER within the categorical model has been widely acknowledged and researched, the definition of the prototypical expressions covers only a small portion of specific categories and cannot capture the full repertoire of expressive behaviors for realistic interactions. Two additional models were developed to describe a larger range of emotional landscapes: the FACS model [10], [137], where various facial muscle AUs were combined to describe the visible appearance changes in facial expressions, and the dimensional model [11], [214], where two continuous-valued variables, namely, valence and arousal, are proposed to continuously encode small changes in the intensity of emotions. Another novel definition, i.e., compound expression, was proposed by Du et al. [51], who argued that several facial emotions are actually combinations of more than one basic expression. These works improve the characterization of facial expressions and, to some extent, can complement the categorical model.

For instance, as discussed above, the visualization results of CNNs have demonstrated a certain congruity between the learned representations and the facial areas defined by AUs. Thus, we can design deep neural network filters to distribute different weights according to the importance degree of different facial muscle action parts. Also, combinations of dimensional models of affect will become even more relevant as a more natural way of dealing with continuous data. Another current direction from deep learning research is the visual attention-based networks that can highlight the most relevant AU-related regions to the FER task through attention mechanisms and allow models to learn expression-discriminative representations.

6.4 Multimodal affect recognition

Finally, human expressive behaviors in realistic applications involve encoding from different perspectives, and the facial expression is only one modality. With the advancement of social media and user-generated content, a large amount of data is uploaded by the users from various platforms, such as text (e.g., Twitter and Facebook), image (e.g., Flickr and Instagram), audio (e.g., podcasts) and video (e.g., YouTube). And multimodal sentiment analysis has become increasingly popular in processing these diverse modalities and analyzing human's opinion (usually, positive or negative) towards a certain entity [215], [216].

To combine useful information from different modalities, recent multimodal sentiment analysis approaches focus on deep neural networks and propose different multi-sensor data fusion methods. Generally, the fusion methods can be categorized into decision-level fusion and feature-level fusion [217]. In decision-level fusion, results from different models are aggregated together at a later stage. In feature-level fusion, features are extracted from each modality independently at an early stage and then combined jointly for a complete representation. For example, CNN with multiple kernel learning (MKL) [218], [219] is employed to fuse acoustic, visual and textual features. Other related research proposes exploring the interactions between different modalities and can achieve better performances in multimodal analysis. For instance, [220] used word-level modality fusion to align each word to corresponding video frames and audio segments. And [221] proposed a tensor fusion network to model both the intra-modality and inter-modality dynamics. More recently, [222] projected the extracted features from each modality to a four-dimensional AffectiveSpace and use a convolutional fuzzy sentiment classifier to predict the degree of a particular emotion in AffectiveSpace. Hence, complex partial emotions can be visualized in a low computational complexity.

Additionally, the fusion of other modalities, such as infrared images, depth information from 3D face models and physiological data, is becoming a promising research direction due to the large complementarity for facial expressions and the beneficial application value for human-computer interaction (HCI) applications.

REFERENCES

[1] C. Darwin and P. Prodger, *The expression of the emotions in man and animals*. Oxford University Press, USA, 1998.

[2] Y.-I. Tian, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 23, no. 2, pp. 97–115, 2001.

[3] P. Ekman and W. V. Friesen, "Constants across cultures in the face and emotion." *Journal of personality and social psychology*, vol. 17, no. 2, pp. 124–129, 1971.

[4] P. Ekman, "Strong evidence for universals in facial expressions: a reply to russell's mistaken critique," *Psychological bulletin*, vol. 115, no. 2, pp. 268–287, 1994.

[5] D. Matsumoto, "More evidence for the universality of a contempt expression," *Motivation and Emotion*, vol. 16, no. 4, pp. 363–368, 1992.

[6] R. E. Jack, O. G. Garrod, H. Yu, R. Caldara, and P. G. Schyns, "Facial expressions of emotion are not culturally universal," *Proceedings of the National Academy of Sciences*, vol. 109, no. 19, pp. 7241–7244, 2012.

[7] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009.

[8] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic analysis of facial affect: A survey of registration, representation, and recognition," *IEEE transactions on pattern analysis and machine intelligence*, vol. 37, no. 6, pp. 1113–1133, 2015.

[9] B. Martinez and M. F. Valstar, "Advances, challenges, and opportunities in automatic facial expression recognition," in *Advances in Face Detection and Facial Image Analysis*. Springer, 2016, pp. 63–100.

[10] P. Ekman, "Facial action coding system (facs)," *A human face*, 2002.

[11] H. Gunes and B. Schuller, "Categorical and dimensional affect analysis in continuous input: Current trends and future directions," *Image and Vision Computing*, vol. 31, no. 2, pp. 120–136, 2013.

[12] 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.

[13] P. Liu, S. Han, Z. Meng, and Y. Tong, "Facial expression recognition via a boosted deep belief network," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1805–1812.

[14] A. Mollahosseini, D. Chan, and M. H. Mahoor, "Going deeper in facial expression recognition using deep neural networks," in *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*. IEEE, 2016, pp. 1–10.

[15] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 29, no. 6, pp. 915–928, 2007.

[16] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 2983–2991.

[17] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos, and S. Yan, "Peak-piloted deep network for facial expression recognition," in *European conference on computer vision*. Springer, 2016, pp. 425–442.

[18] 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 transactions on pattern analysis and machine intelligence*, vol. 38, no. 8, pp. 1548–1568, 2016.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

[19] R. Zhi, M. Flierl, Q. Ruan, and W. B. Kleijn, "Graph-preserving sparse nonnegative matrix factorization with application to facial expression recognition," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 41, no. 1, pp. 38–52, 2011.

[20] L. Zhong, Q. Liu, P. Yang, B. Liu, J. Huang, and D. N. Metaxas, "Learning active facial patches for expression analysis," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2562–2569.

[21] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee *et al.*, "Challenges in representation learning: A report on three machine learning contests," in *International Conference on Neural Information Processing*. Springer, 2013, pp. 117–124.

[22] A. Dhall, O. Ramana Murthy, R. Goecke, J. Joshi, and T. Gedeon, "Video and image based emotion recognition challenges in the wild: Emotiw 2015," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 423–426.

[23] A. Dhall, R. Goecke, S. Ghosh, J. Joshi, J. Hoey, and T. Gedeon, "From individual to group-level emotion recognition: Emotiw 5.0," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 524–528.

[24] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[25] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[26] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.

[28] M. Pantic and L. J. M. Rothkrantz, "Automatic analysis of facial expressions: The state of the art," *IEEE Transactions on pattern analysis and machine intelligence*, vol. 22, no. 12, pp. 1424–1445, 2000.

[29] B. Fasel and J. Luettin, "Automatic facial expression analysis: a survey," *Pattern recognition*, vol. 36, no. 1, pp. 259–275, 2003.

[30] P. V. Rouast, M. Adam, and R. Chiong, "Deep learning for human affect recognition: Insights and new developments," *IEEE Transactions on Affective Computing*, pp. 1–1, 2019.

[31] M. F. Valstar, M. Mehu, B. Jiang, M. Pantic, and K. Scherer, "Meta-analysis of the first facial expression recognition challenge," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 42, no. 4, pp. 966–979, 2012.

[32] 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 *Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on*. IEEE, 2010, pp. 94–101.

[33] G. Zhao, X. Huang, M. Taini, S. Z. Li, and M. PietikäInen, "Facial expression recognition from near-infrared videos," *Image and Vision Computing*, vol. 29, no. 9, pp. 607–619, 2011.

[34] S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2017, pp. 2584–2593.

[35] 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, p. 65.

[36] M. Lyons, S. Akamatsu, M. Kamachi, and J. Gyoba, "Coding facial expressions with gabor wavelets," in *Automatic Face and Gesture Recognition, 1998. Proceedings. Third IEEE International Conference on*. IEEE, 1998, pp. 200–205.

[37] J. M. Susskind, A. K. Anderson, and G. E. Hinton, "The toronto face database," *Department of Computer Science, University of Toronto, Toronto, ON, Canada, Tech. Rep*, vol. 3, 2010.

[38] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[39] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic Face and Gesture Recognition, 2006. 7th International Conference on*. IEEE, 2006, pp. 211–216.

[40] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *The 8th International Conference on Automatic Face and Gesture Recognition. Amsterdam, The Netherlands*. IEEE, 2008.

[41] O. Langner, R. Dotsch, G. Bijlstra, D. H. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and validation of the radboud faces database," *Cognition and Emotion*, vol. 24, no. 8, pp. 1377–1388, 2010.

[42] D. Lundqvist, A. Flykt, and A. Öhman, "The karolinska directed emotional faces (kdef)," *CD ROM from Department of Clinical Neuroscience, Psychology section, Karolinska Institutet*, no. 1998, 1998.

[43] C. F. Benitez-Quiroz, R. Srinivasan, and A. M. Martinez, "Emotionet: An accurate, real-time algorithm for the automatic annotation of a million facial expressions in the wild," in *Proceedings of IEEE International Conference on Computer Vision & Pattern Recognition (CVPR), Las Vegas, NV, USA*, 2016.

[44] S. Li and W. Deng, "Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition," *IEEE Transactions on Image Processing*, vol. 28, no. 1, pp. 356–370, Jan 2019.

[45] A. Mollahosseini, B. Hasani, and M. H. Mahoor, "Affectnet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Transactions on Affective Computing*, vol. PP, no. 99, pp. 1–1, 2017.

[46] Z. Zhang, P. Luo, C. L. Chen, and X. Tang, "From facial expression recognition to interpersonal relation prediction," *International Journal of Computer Vision*, vol. 126, no. 5, pp. 1–20, 2018.

[47] S. Cheng, I. Kotsia, M. Pantic, and S. Zafeiriou, "4dfab: A large scale 4d database for facial expression analysis and biometric applications," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2018, pp. 5117–5126.

[48] A. Dhall, R. Goecke, S. Lucey, T. Gedeon *et al.*, "Collecting large, richly annotated facial-expression databases from movies," *IEEE multimedia*, vol. 19, no. 3, pp. 34–41, 2012.

[49] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Static facial expression analysis in tough conditions: Data, evaluation protocol and benchmark," in *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*. IEEE, 2011, pp. 2106–2112.

[50] C. F. Benitez-Quiroz, R. Srinivasan, Q. Feng, Y. Wang, and A. M. Martinez, "Emotionet challenge: Recognition of facial expressions of emotion in the wild," *arXiv preprint arXiv:1703.01210*, 2017.

[51] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proceedings of the National Academy of Sciences*, vol. 111, no. 15, pp. E1454–E1462, 2014.

[52] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, vol. 1. IEEE, 2001, pp. I–I.

[53] Z. Yu and C. Zhang, "Image based static facial expression recognition with multiple deep network learning," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 435–442.

[54] B.-K. Kim, H. Lee, J. Roh, and S.-Y. Lee, "Hierarchical committee of deep cnns with exponentially-weighted decision fusion for static facial expression recognition," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 427–434.

[55] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*. IEEE, 2012, pp. 2879–2886.

[56] F. De la Torre, W.-S. Chu, X. Xiong, F. Vicente, X. Ding, and J. F. Cohn, "Intraface," in *IEEE International Conference on Automatic Face and Gesture Recognition (FG)*, 2015.

[57] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, no. 6, pp. 681–685, 2001.

[58] N. Zeng, H. Zhang, B. Song, W. Liu, Y. Li, and A. M. Dobaie, "Facial expression recognition via learning deep sparse autoencoders," *Neurocomputing*, vol. 273, pp. 643–649, 2018.

[59] B. Hasani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 790–795.

[60] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550.

[61] T. Devries, K. Biswaranjan, and G. W. Taylor, "Multi-task learning of facial landmarks and expression," in *Computer and Robot Vision (CRV), 2014 Canadian Conference on*. IEEE, 2014, pp. 98–103.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

[62] B. Sun, L. Li, G. Zhou, and J. He, "Facial expression recognition in the wild based on multimodal texture features," *Journal of Electronic Imaging*, vol. 25, no. 6, p. 061407, 2016.

[63] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3444–3451.

[64] M. Shin, M. Kim, and D.-S. Kwon, "Baseline cnn structure analysis for facial expression recognition," in *Robot and Human Interactive Communication (RO-MAN), 2016 25th IEEE International Symposium on*. IEEE, 2016, pp. 724–729.

[65] Z. Meng, P. Liu, J. Cai, S. Han, and Y. Tong, "Identity-aware convolutional neural network for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 558–565.

[66] X. Xiong and F. De la Torre, "Supervised descent method and its applications to face alignment," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 532–539.

[67] H.-W. Ng, V. D. Nguyen, V. Vonikakis, and S. Winkler, "Deep learning for emotion recognition on small datasets using transfer learning," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 443–449.

[68] H. Ding, S. K. Zhou, and R. Chellappa, "Facenet2expnet: Regularizing a deep face recognition net for expression recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 118–126.

[69] S. Ren, X. Cao, Y. Wei, and J. Sun, "Face alignment at 3000 fps via regressing local binary features," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2014, pp. 1685–1692.

[70] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Incremental face alignment in the wild," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 1859–1866.

[71] D. H. Kim, W. Baddar, J. Jang, and Y. M. Ro, "Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 223–236, 2019.

[72] Y. Sun, X. Wang, and X. Tang, "Deep convolutional network cascade for facial point detection," in *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*. IEEE, 2013, pp. 3476–3483.

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

[74] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[75] Z. Yu, Q. Liu, and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer*, pp. 1–9, 2017.

[76] X. Ouyang, S. Kawaai, E. G. H. Goh, S. Shen, W. Ding, H. Ming, and D.-Y. Huang, "Audio-visual emotion recognition using deep transfer learning and multiple temporal models," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 577–582.

[77] Z. Yu, G. Liu, Q. Liu, and J. Deng, "Spatio-temporal convolutional features with nested lstm for facial expression recognition," *Neurocomputing*, vol. 317, pp. 50–57, 2018.

[78] X. Liu, B. Kumar, J. You, and P. Jia, "Adaptive deep metric learning for identity-aware facial expression recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops (CVPRW)*, 2017, pp. 522–531.

[79] G. Levi and T. Hassner, "Emotion recognition in the wild via convolutional neural networks and mapped binary patterns," in *Proceedings of the 2015 ACM on international conference on multimodal interaction*. ACM, 2015, pp. 503–510.

[80] W. Li, M. Li, Z. Su, and Z. Zhu, "A deep-learning approach to facial expression recognition with candid images," in *Machine Vision Applications (MVA), 2015 14th IAPR International Conference on*. IEEE, 2015, pp. 279–282.

[81] I. Abbasnejad, S. Sridharan, D. Nguyen, S. Denman, C. Fookes, and S. Lucey, "Using synthetic data to improve facial expression analysis with 3d convolutional networks," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 1609–1618.

[82] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," in *Advances in neural information processing systems*, 2014, pp. 2672–2680.

[83] J. Li and E. Y. Lam, "Facial expression recognition using deep neural networks," in *Imaging Systems and Techniques (IST), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1–6.

[84] S. Ebrahimi Kahou, V. Michalski, K. Konda, R. Memisevic, and C. Pal, "Recurrent neural networks for emotion recognition in video," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*. ACM, 2015, pp. 467–474.

[85] S. A. Bargal, E. Barsoum, C. C. Ferrer, and C. Zhang, "Emotion recognition in the wild from videos using images," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 433–436.

[86] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, "Enhancing cnn with preprocessing stage in automatic emotion recognition," *Procedia Computer Science*, vol. 116, pp. 523–529, 2017.

[87] C.-M. Kuo, S.-H. Lai, and M. Sarkis, "A compact deep learning model for robust facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 2121–2129.

[88] A. Yao, D. Cai, P. Hu, S. Wang, L. Sha, and Y. Chen, "Holonet: towards robust emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 472–478.

[89] P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 553–560.

[90] T. Hassner, S. Harel, E. Paz, and R. Enbar, "Effective face frontalization in unconstrained images," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 4295–4304.

[91] C. Sagonas, Y. Panagakis, S. Zafeiriou, and M. Pantic, "Robust statistical face frontalization," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3871–3879.

[92] L. Deng, D. Yu *et al.*, "Deep learning: methods and applications," *Foundations and Trends® in Signal Processing*, vol. 7, no. 3–4, pp. 197–387, 2014.

[93] Y. Tang, "Deep learning using linear support vector machines," *arXiv preprint arXiv:1306.0239*, 2013.

[94] A. Dapogny and K. Bailly, "Investigating deep neural forests for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 629–633.

[95] P. Kontschieder, M. Fiterau, A. Criminisi, and S. Rota Bulo, "Deep neural decision forests," in *Proceedings of the IEEE international conference on computer vision*, 2015, pp. 1467–1475.

[96] J. Donahue, Y. Jia, O. Vinyals, J. Hoffman, N. Zhang, E. Tzeng, and T. Darrell, "Decaf: A deep convolutional activation feature for generic visual recognition," in *International conference on machine learning*, 2014, pp. 647–655.

[97] A. S. Razavian, H. Azizpour, J. Sullivan, and S. Carlsson, "Cnn features off-the-shelf: an astounding baseline for recognition," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*. IEEE, 2014, pp. 512–519.

[98] N. Otberdout, A. Kacem, M. Daoudi, L. Ballihi, and S. Berretti, "Deep covariance descriptors for facial expression recognition," in *BMVC*, 2018.

[99] D. Acharya, Z. Huang, D. Pani Paudel, and L. Van Gool, "Covariance pooling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 367–374.

[100] S. Ouellet, "Real-time emotion recognition for gaming using deep convolutional network features," *arXiv preprint arXiv:1408.3750*, 2014.

[101] M. Liu, S. Li, S. Shan, and X. Chen, "Au-aware deep networks for facial expression recognition," in *Automatic Face and Gesture Recognition (FG), 2013 10th IEEE International Conference and Workshops on*. IEEE, 2013, pp. 1–6.

[102] ——, "Au-inspired deep networks for facial expression feature learning," *Neurocomputing*, vol. 159, pp. 126–136, 2015.

[103] P. Khorrami, T. Paine, and T. Huang, "Do deep neural networks learn facial action units when doing expression recognition?" *arXiv preprint arXiv:1510.02969v3*, 2015.

[104] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Island loss for learning discriminative features in facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 302–309.

[105] H. Yang, U. Ciftci, and L. Yin, "Facial expression recognition by de-expression residue learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168–2177.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

[106] D. Hamester, P. Barros, and S. Wermter, "Face expression recognition with a 2-channel convolutional neural network," in *Neural Networks (IJCNN), 2015 International Joint Conference on*. IEEE, 2015, pp. 1–8.

[107] X. Liu, B. V. Kumar, P. Jia, and J. You, "Hard negative generation for identity-disentangled facial expression recognition," *Pattern Recognition*, vol. 88, pp. 1 – 12, 2019.

[108] S. Reed, K. Sohn, Y. Zhang, and H. Lee, "Learning to disentangle factors of variation with manifold interaction," in *International Conference on Machine Learning*, 2014, pp. 1431–1439.

[109] Z. Zhang, P. Luo, C.-C. Loy, and X. Tang, "Learning social relation traits from face images," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 3631–3639.

[110] Y. Guo, D. Tao, J. Yu, H. Xiong, Y. Li, and D. Tao, "Deep neural networks with relativity learning for facial expression recognition," in *Multimedia & Expo Workshops (ICMEW), 2016 IEEE International Conference on*. IEEE, 2016, pp. 1–6.

[111] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim, and S.-Y. Lee, "Fusing aligned and non-aligned face information for automatic affect recognition in the wild: A deep learning approach," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2016, pp. 48–57.

[112] C. Pramerdorfer and M. Kampel, "Facial expression recognition using convolutional neural networks: State of the art," *arXiv preprint arXiv:1612.02903*, 2016.

[113] M.-I. Georgescu, R. T. Ionescu, and M. Popescu, "Local learning with deep and handcrafted features for facial expression recognition," *IEEE Access*, vol. 7, pp. 64 827–64 836, 2019.

[114] O. M. Parkhi, A. Vedaldi, A. Zisserman *et al.*, "Deep face recognition." in *BMVC*, vol. 1, no. 3, 2015, p. 6.

[115] T. Kaneko, K. Hiramatsu, and K. Kashino, "Adaptive visual feedback generation for facial expression improvement with multi-task deep neural networks," in *Proceedings of the 2016 ACM on Multimedia Conference*. ACM, 2016, pp. 327–331.

[116] H. Kaya, F. Gürpınar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66–75, 2017.

[117] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko, "Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video," *arXiv preprint arXiv:1711.04598*, 2017.

[118] D. G. Lowe, "Object recognition from local scale-invariant features," in *Computer vision, 1999. The proceedings of the seventh IEEE international conference on*, vol. 2. Ieee, 1999, pp. 1150–1157.

[119] 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 Transactions on Multimedia*, vol. 18, no. 12, pp. 2528–2536, 2016.

[120] Z. Luo, J. Chen, T. Takiguchi, and Y. Ariki, "Facial expression recognition with deep age," in *Multimedia & Expo Workshops (ICMEW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 657–662.

[121] L. Chen, M. Zhou, W. Su, M. Wu, J. She, and K. Hirota, "Softmax regression based deep sparse autoencoder network for facial emotion recognition in human-robot interaction," *Information Sciences*, vol. 428, pp. 49–61, 2018.

[122] V. Mavani, S. Raman, and K. P. Miyapuram, "Facial expression recognition using visual saliency and deep learning," *arXiv preprint arXiv:1708.08016*, 2017.

[123] B.-F. Wu and C.-H. Lin, "Adaptive feature mapping for customizing deep learning based facial expression recognition model," *IEEE Access*, 2018.

[124] Y. Liu, X. Yuan, X. Gong, Z. Xie, F. Fang, and Z. Luo, "Conditional convolution neural network enhanced random forest for facial expression recognition," *Pattern Recognition*, vol. 84, pp. 251 – 261, 2018.

[125] W. Sun, H. Zhao, and Z. Jin, "A visual attention based roi detection method for facial expression recognition," *Neurocomputing*, vol. 296, pp. 12 – 22, 2018.

[126] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion aware facial expression recognition using cnn with attention mechanism," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2439–2450, 2019.

[127] S. Xie, H. Hu, and Y. Wu, "Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition," *Pattern Recognition*, vol. 92, pp. 177 – 191, 2019.

[128] W. Shang, K. Sohn, D. Almeida, and H. Lee, "Understanding and improving convolutional neural networks via concatenated rectified linear units," in *International Conference on Machine Learning*, 2016, pp. 2217–2225.

[129] J. Zeng, S. Shan, and X. Chen, "Facial expression recognition with inconsistently annotated datasets," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 222–237.

[130] Y. Wen, K. Zhang, Z. Li, and Y. Qiao, "A discriminative feature learning approach for deep face recognition," in *European Conference on Computer Vision*. Springer, 2016, pp. 499–515.

[131] F. Schroff, D. Kalenichenko, and J. Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 815–823.

[132] D. Ciregan, U. Meier, and J. Schmidhuber, "Multi-column deep neural networks for image classification," in *Computer vision and pattern recognition (CVPR), 2012 IEEE conference on*. IEEE, 2012, pp. 3642–3649.

[133] G. Pons and D. Masip, "Supervised committee of convolutional neural networks in automated facial expression analysis," *IEEE Transactions on Affective Computing*, vol. 9, no. 3, pp. 343–350, 2018.

[134] K. Liu, M. Zhang, and Z. Pan, "Facial expression recognition with cnn ensemble," in *Cyberworlds (CW), 2016 International Conference on*. IEEE, 2016, pp. 163–166.

[135] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 423–430.

[136] G. Pons and D. Masip, "Multi-task, multi-label and multi-domain learning with residual convolutional networks for emotion recognition," *arXiv preprint arXiv:1802.06664*, 2018.

[137] P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[138] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa, "An all-in-one convolutional neural network for face analysis," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 17–24.

[139] Y. Jang, H. Gunes, and I. Patras, "Smilenet: registration-free smiling face detection in the wild," in *Proceedings of the IEEE International Conference on Computer Vision*, 2017, pp. 1581–1589.

[140] Y. Lv, Z. Feng, and C. Xu, "Facial expression recognition via deep learning," in *Smart Computing (SMARTCOMP), 2014 International Conference on*. IEEE, 2014, pp. 303–308.

[141] S. Rifai, Y. Bengio, A. Courville, P. Vincent, and M. Mirza, "Disentangling factors of variation for facial expression recognition," in *European Conference on Computer Vision*. Springer, 2012, pp. 808–822.

[142] Y.-H. Lai and S.-H. Lai, "Emotion-preserving representation learning via generative adversarial network for multi-view facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 263–270.

[143] F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Joint pose and expression modeling for facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018, pp. 3359–3368.

[144] H. Yang, Z. Zhang, and L. Yin, "Identity-adaptive facial expression recognition through expression regeneration using conditional generative adversarial networks," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 294–301.

[145] J. Chen, J. Konrad, and P. Ishwar, "Vgan-based image representation learning for privacy-preserving facial expression recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018, pp. 1570–1579.

[146] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean, P. Froumenty, Y. Dauphin, N. Boulanger-Lewandowski *et al.*, "Emonets: Multimodal deep learning approaches for emotion recognition in video," *Journal on Multimodal User Interfaces*, vol. 10, no. 2, pp. 99–111, 2016.

[147] M. Liu, R. Wang, S. Li, S. Shan, Z. Huang, and X. Chen, "Combining multiple kernel methods on riemannian manifold for emotion recognition in the wild," in *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 2014, pp. 494–501.

[148] W. Ding, M. Xu, D. Huang, W. Lin, M. Dong, X. Yu, and H. Li, "Audio and face video emotion recognition in the wild using deep neural networks and small datasets," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 506–513.

[149] B. Xu, Y. Fu, Y.-G. Jiang, B. Li, and L. Sigal, "Video emotion recognition with transferred deep feature encodings," in *Proceedings*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

of the 2016 ACM on International Conference on Multimedia Retrieval. ACM, 2016, pp. 15–22.

[150] Y. Kim, B. Yoo, Y. Kwak, C. Choi, and J. Kim, "Deep generative-contrastive networks for facial expression recognition," *arXiv preprint arXiv:1703.07140*, 2017.

[151] N. Sun, Q. Li, R. Huan, J. Liu, and G. Han, "Deep spatial-temporal feature fusion for facial expression recognition in static images," *Pattern Recognition Letters*, 2017.

[152] S. Kumawat, M. Verma, and S. Raman, "Lbvcnn: Local binary volume convolutional neural network for facial expression recognition from image sequences," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[153] B. Hasani and M. H. Mahoor, "Facial expression recognition using enhanced deep 3d convolutional neural networks," in *Computer Vision and Pattern Recognition Workshops (CVPRW), 2017 IEEE Conference on*. IEEE, 2017, pp. 2278–2288.

[154] S. Wang, Z. Zheng, S. Yin, J. Yang, and Q. Ji, "A novel dynamic model capturing spatial and temporal patterns for facial expression analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2019.

[155] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang, Y. Zong, and N. Sun, "Multi-clue fusion for emotion recognition in the wild," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 458–463.

[156] Z. Cui, S. Xiao, Z. Niu, S. Yan, and W. Zheng, "Recurrent shape regression," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2018.

[157] Y. Fan, X. Lu, D. Li, and Y. Liu, "Video-based emotion recognition using cnn-rnn and c3d hybrid networks," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 445–450.

[158] V. Vielzeuf, S. Pateux, and F. Jurie, "Temporal multimodal fusion for video emotion classification in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 569–576.

[159] J. Chen, R. Xu, and L. Liu, "Deep peak-neutral difference feature for facial expression recognition," *Multimedia Tools and Applications*, pp. 1–17, 2018.

[160] Q. V. Le, N. Jaitly, and G. E. Hinton, "A simple way to initialize recurrent networks of rectified linear units," *arXiv preprint arXiv:1504.00941*, 2015.

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

[162] D. A. AL CHANTI and A. Caplier, "Deep learning for spatio-temporal modeling of dynamic spontaneous emotions," *IEEE Transactions on Affective Computing*, pp. 1–1, 2018.

[163] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Computer Vision (ICCV), 2015 IEEE International Conference on*. IEEE, 2015, pp. 4489–4497.

[164] P. Barros and S. Wermter, "Developing crossmodal expression recognition based on a deep neural model," *Adaptive behavior*, vol. 24, no. 5, pp. 373–396, 2016.

[165] J. Zhao, X. Mao, and J. Zhang, "Learning deep facial expression features from image and optical flow sequences using 3d cnn," *The Visual Computer*, pp. 1–15, 2018.

[166] 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*. Springer, 2014, pp. 143–157.

[167] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan, "Object detection with discriminatively trained part-based models," *IEEE transactions on pattern analysis and machine intelligence*, vol. 32, no. 9, pp. 1627–1645, 2010.

[168] D. Nguyen, K. Nguyen, S. Sridharan, A. Ghasemi, D. Dean, and C. Fookes, "Deep spatio-temporal features for multimodal emotion recognition," in *Applications of Computer Vision (WACV), 2017 IEEE Winter Conference on*. IEEE, 2017, pp. 1215–1223.

[169] S. Pini, O. B. Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 536–543.

[170] R. Arandjelovic, P. Gronat, A. Torii, T. Pajdla, and J. Sivic, "Netvlad: Cnn architecture for weakly supervised place recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 5297–5307.

[171] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song, "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*. ACM, 2017, pp. 529–535.

[172] 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 *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 2625–2634.

[173] D. K. Jain, Z. Zhang, and K. Huang, "Multi angle optimal pattern-based deep learning for automatic facial expression recognition," *Pattern Recognition Letters*, 2017.

[174] S. Kankanamge, C. Fookes, and S. Sridharan, "Facial analysis in the wild with lstm networks," in *Image Processing (ICIP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 1052–1056.

[175] W. J. Baddar and Y. M. Ro, "Learning spatio-temporal features with partial expression sequences for on-the-fly prediction," in *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[176] M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt, "Spatio-temporal convolutional sparse auto-encoder for sequence classification." in *BMVC*, 2012, pp. 1–12.

[177] K. Simonyan and A. Zisserman, "Two-stream convolutional networks for action recognition in videos," in *Advances in neural information processing systems*, 2014, pp. 568–576.

[178] J. Susskind, V. Mnih, G. Hinton *et al.*, "On deep generative models with applications to recognition," in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*. IEEE, 2011, pp. 2857–2864.

[179] Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP), 2014 Tenth International Conference on*. IEEE, 2014, pp. 211–214.

[180] M. Xu, W. Cheng, Q. Zhao, L. Ma, and F. Xu, "Facial expression recognition based on transfer learning from deep convolutional networks," in *Natural Computation (ICNC), 2015 11th International Conference on*. IEEE, 2015, pp. 702–708.

[181] Y. Liu, J. Zeng, S. Shan, and Z. Zheng, "Multi-channel pose-aware convolution neural networks for multi-view facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 458–465.

[182] E. P. Ijjina and C. K. Mohan, "Facial expression recognition using kinect depth sensor and convolutional neural networks," in *Machine Learning and Applications (ICMLA), 2014 13th International Conference on*. IEEE, 2014, pp. 392–396.

[183] M. Z. Uddin, M. M. Hassan, A. Almogren, M. Zuair, G. Fortino, and J. Torresen, "A facial expression recognition system using robust face features from depth videos and deep learning," *Computers & Electrical Engineering*, vol. 63, pp. 114–125, 2017.

[184] W. Li, D. Huang, H. Li, and Y. Wang, "Automatic 4d facial expression recognition using dynamic geometrical image network," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 24–30.

[185] F.-J. Chang, A. T. Tran, T. Hassner, I. Masi, R. Nevatia, and G. Medioni, "Expnet: Landmark-free, deep, 3d facial expressions," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 122–129.

[186] Z. Chen, D. Huang, Y. Wang, and L. Chen, "Fast and light manifold cnn based 3d facial expression recognition across pose variations," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 229–238.

[187] O. K. Oyedotun, G. Demisse, A. E. R. Shabayek, D. Aouada, and B. Ottersten, "Facial expression recognition via joint deep learning of rgb-depth map latent representations," in *2017 IEEE International Conference on Computer Vision Workshop (ICCVW)*, 2017.

[188] H. Li, J. Sun, Z. Xu, and L. Chen, "Multimodal 2d+ 3d facial expression recognition with deep fusion convolutional neural network," *IEEE Transactions on Multimedia*, vol. 19, no. 12, pp. 2816–2831, 2017.

[189] A. Jan, H. Ding, H. Meng, L. Chen, and H. Li, "Accurate facial parts localization and deep learning for 3d facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 466–472.

[190] J. M. Susskind, G. E. Hinton, J. R. Movellan, and A. K. Anderson, "Generating facial expressions with deep belief nets," in *Affective Computing*. InTech, 2008.

[191] M. Sabzevari, S. Toosizadeh, S. R. Quchani, and V. Abrishami, "A fast and accurate facial expression synthesis system for color face images using face graph and deep belief network," in *Electronics and*

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.

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/TAFFC.2020.2981446, IEEE Transactions on Affective Computing

Information Engineering (ICEIE), 2010 International Conference On, vol. 2. IEEE, 2010, pp. V2–354.

[192] V. Mnih, J. M. Susskind, G. E. Hinton *et al.*, "Modeling natural images using gated mrfs," *IEEE transactions on pattern analysis and machine intelligence*, vol. 35, no. 9, pp. 2206–2222, 2013.

[193] R. Yeh, Z. Liu, D. B. Goldman, and A. Agarwala, "Semantic facial expression editing using autoencoded flow," *arXiv preprint arXiv:1611.09961*, 2016.

[194] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *AAAI*, 2018, p. 6781–6788.

[195] L. Song, Z. Lu, R. He, Z. Sun, and T. Tan, "Geometry guided adversarial facial expression synthesis," in *2018 ACM Multimedia Conference on Multimedia Conference*. ACM, 2018, pp. 627–635.

[196] A. Pumarola, A. Agudo, A. M. Martinez, A. Sanfeliu, and F. Moreno-Noguer, "Ganimation: Anatomically-aware facial animation from a single image," in *Proceedings of the European Conference on Computer Vision (ECCV)*, 2018, pp. 818–833.

[197] F. Qiao, N. Yao, Z. Jiao, Z. Li, H. Chen, and H. Wang, "Geometry-contrastive generative adversarial network for facial expression synthesis," *arXiv preprint arXiv:1802.01822*, 2018.

[198] Z. Geng, C. Cao, and S. Tulyakov, "3d guided fine-grained face manipulation," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[199] I. Masi, A. T. Tran, T. Hassner, J. T. Leksut, and G. Medioni, "Do we really need to collect millions of faces for effective face recognition?" in *European Conference on Computer Vision*. Springer, 2016, pp. 579–596.

[200] N. Mousavi, H. Siqueira, P. Barros, B. Fernandes, and S. Wermter, "Understanding how deep neural networks learn face expressions," in *Neural Networks (IJCNN), 2016 International Joint Conference on*. IEEE, 2016, pp. 227–234.

[201] R. Breuer and R. Kimmel, "A deep learning perspective on the origin of facial expressions," *arXiv preprint arXiv:1705.01842*, 2017.

[202] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

[203] I. Lüsi, J. C. J. Junior, J. Gorbova, X. Baró, S. Escalera, H. Demirel, J. Allik, C. Ozcinar, and G. Anbarjafari, "Joint challenge on dominant and complementary emotion recognition using micro emotion features and head-pose estimation: Databases," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 809–813.

[204] J. Wan, S. Escalera, X. Baro, H. J. Escalante, I. Guyon, M. Madadi, J. Allik, J. Gorbova, and G. Anbarjafari, "Results and analysis of chalearn lap multi-modal isolated and continuous gesture recognition, and real versus fake expressed emotions challenges," in *ChaLearn LaP, Action, Gesture, and Emotion Recognition Workshop and Competitions: Large Scale Multimodal Gesture Recognition and Real versus Fake expressed emotions, ICCV*, vol. 4, no. 6, 2017.

[205] Y.-G. Kim and X.-P. Huynh, "Discrimination between genuine versus fake emotion using long-short term memory with parametric bias and facial landmarks," in *Computer Vision Workshop (ICCVW), 2017 IEEE International Conference on*. IEEE, 2017, pp. 3065–3072.

[206] L. Li, T. Baltrusaitis, B. Sun, and L.-P. Morency, "Combining sequential geometry and texture features for distinguishing genuine and deceptive emotions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017, pp. 3147–3153.

[207] J. Guo, S. Zhou, J. Wu, J. Wan, X. Zhu, Z. Lei, and S. Z. Li, "Multi-modality network with visual and geometrical information for micro emotion recognition," in *Automatic Face & Gesture Recognition (FG 2017), 2017 12th IEEE International Conference on*. IEEE, 2017, pp. 814–819.

[208] R. Vemulapalli and A. Agarwala, "A compact embedding for facial expression similarity," in *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2019.

[209] E. Barsoum, C. Zhang, C. C. Ferrer, and Z. Zhang, "Training deep networks for facial expression recognition with crowd-sourced label distribution," in *Proceedings of the 18th ACM International Conference on Multimodal Interaction*. ACM, 2016, pp. 279–283.

[210] "Chapter 19 - affective facial computing: Generalizability across domains," in *Multimodal Behavior Analysis in the Wild*, X. Alameda-Pineda, E. Ricci, and N. Sebe, Eds. Academic Press, 2019, pp. 407 – 441.

[211] X. Wei, H. Li, J. Sun, and L. Chen, "Unsupervised domain adaptation with regularized optimal transport for multimodal 2d+ 3d facial expression recognition," in *Automatic Face & Gesture Recognition (FG 2018), 2018 13th IEEE International Conference on*. IEEE, 2018, pp. 31–37.

[212] S. Li and W. Deng, "Deep emotion transfer network for cross-database facial expression recognition," in *Pattern Recognition (ICPR), 2018 26th International Conference*. IEEE, 2018, pp. 3092–3099.

[213] R. L. Testa, C. G. Corrêa, A. Machado-Lima, and F. L. S. Nunes, "Synthesis of facial expressions in photographs: Characteristics, approaches, and challenges," *ACM Comput. Surv.*, vol. 51, no. 6, pp. 124:1–124:35, 2019.

[214] J. A. Russell, "A circumplex model of affect." *Journal of personality and social psychology*, vol. 39, no. 6, p. 1161, 1980.

[215] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, and M. Pantic, "A survey of multimodal sentiment analysis," *Image and Vision Computing*, vol. 65, pp. 3–14, 2017.

[216] I. Chaturvedi, E. Cambria, R. E. Welsch, and F. Herrera, "Distinguishing between facts and opinions for sentiment analysis: Survey and challenges," *Information Fusion*, vol. 44, pp. 65–77, 2018.

[217] S. Poria, E. Cambria, and A. Gelbukh, "Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis," in *Proceedings of the 2015 conference on empirical methods in natural language processing*, 2015, pp. 2539–2544.

[218] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, "Convolutional mkl based multimodal emotion recognition and sentiment analysis," in *2016 IEEE 16th international conference on data mining (ICDM)*. IEEE, 2016, pp. 439–448.

[219] S. Poria, H. Peng, A. Hussain, N. Howard, and E. Cambria, "Ensemble application of convolutional neural networks and multiple kernel learning for multimodal sentiment analysis," *Neurocomputing*, vol. 261, pp. 217–230, 2017.

[220] M. Chen, S. Wang, P. P. Liang, T. Baltrušaitis, A. Zadeh, and L.-P. Morency, "Multimodal sentiment analysis with word-level fusion and reinforcement learning," in *Proceedings of the 19th ACM International Conference on Multimodal Interaction*, 2017, pp. 163–171.

[221] A. Zadeh, M. Chen, S. Poria, E. Cambria, and L.-P. Morency, "Tensor fusion network for multimodal sentiment analysis," *arXiv preprint arXiv:1707.07250*, 2017.

[222] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, "Fuzzy commonsense reasoning for multimodal sentiment analysis," *Pattern Recognition Letters*, vol. 125, no. 264-270, 2019.

Shan Li received her B.E. degree in telecommunication engineering from the Beijing University of Posts and Telecommunications (BUPT), Beijing, China, in 2016. She is currently pursuing a Ph.D. degree in information and telecommunications engineering. Her research interests include facial expression analysis and deep learning.

Weihong Deng received his B.E. degree in information engineering and his Ph.D. degree in signal and information processing from the Beijing University of Posts and Telecommunications (BUPT) in 2004 and 2009, respectively. From Oct. 2007 to Dec. 2008, he was a postgraduate exchange student in the School of Information Technologies, University of Sydney, Australia, under the support of the China Scholarship Council. He is currently a professor in the School of Information and Telecommunications Engineering, BUPT. His research interests include statistical pattern recognition and computer vision.He has published over 100 technical papers in international journals and conferences. He serves as a guest editor for the Image and Vision Computing Journal and a reviewer for several international journals, such as IEEE TPAMI / TIP / TIFS / TNNLS / TMM / TSMC, IJCV, PR / PRL.

1949-3045 (c) 2020 IEEE. Personal use is permitted, but republication/redistribution requires IEEE permission. See http://www.ieee.org/publications_standards/publications/rights/index.html for more information.
Authorized licensed use limited to: BEIJING UNIVERSITY OF POST AND TELECOM. Downloaded on December 14,2021 at 02:52:11 UTC from IEEE Xplore. Restrictions apply.