# Survey on Facial Expression Recognition: History, Applications, and Challenges

Xibin Zhao, Junjie Zhu, Bingjun Luo, and Yue Gao, *Tsinghua University, Beijing, 100084, China*

*Facial expressions are most important channels for transmitting affective information in human interaction. In the early 1970s, due to the limitation of weak computing power and powerless algorithms, automatic facial expression recognition (AFER) entered the freezing period for a long time. However, with the geometrical growth of computing power and the rapid improvement of algorithms, the accuracy of AFER has been significantly enhanced. FER has ushered in the flowering and fruitful application in different fields such as medical treatment, transportation, and business. This article introduces and surveys these recent advances. We first provide a detailed review of the related studies in FER, including emotion representation, well-known datasets, and FER's history. Next, we detailed the background and practical significance of FER technology's application in different fields. We finally summarize some of the scientific and engineering challenges to promote better use of FER in real-world applications.*

Facial expressions (FE) are vital signaling systems of affect, conveying cues about the emotional state of persons. The goal of AFER is to build a robust model, which can automatically encode emotion information from user's facial representations. By understanding emotion information to infer the changes in the user's affective state, many other human–computer interaction systems can be more anticipatory and human centered in the future.

The study of FEs has a long history, began to systematically study the formation mechanism of FEs from the perspectives of muscle performance, psychology, and sociology in the 19th century. Later Charles Darwin also conducted a detailed study on the evolution and development of FEs from the perspective of evolutionary perspective. However, the study of AFER did not break ground until 1978. Trapped by limited computational power and machine learning algorithms, the effect of this method is not ideal, which also led to the fact that AFER did not receive much attention from researchers in the next decade. This dilemma was broken with the birth of the first modern dataset CK[1] in 2000, and AFER entered a new era led by traditional machine learning methods. At the same time, with the continuous improvement of computing power and data,[2,3] the model represented by deep learning has significantly improved the accuracy and robustness of FER. FER has also gained a wide range of applications in more and more fields such as medical treatment, transportation, etc.

A large number of outstanding reviews on FER have been published this year, including traditional methods, deep learning methods, and standard pipelines with fundamental components, namely face registration, representation, dimensionally reduction, and recognition. Nevertheless, they only focus on the review of FER methods, while the application of FE in real scenarios and potential challenges is less mentioned. In this article, we offer a newcomer to this field not only an overall introduction of FER's development, but also the application of FER in real business scenarios, related challenges, and possible solutions.

This article generally reviews on the development of FER, as well as introduces and investigates recent advances in the application of FER. In contrast to previously published survey papers in this field, we focus on the possible difficulties of building a good model that can satisfy the real application requirements, and also offer responding solutions realized in the future. The rest of this article is organized as follows: First section provides a detailed review of

1070-986X © 2021 IEEE
Digital Object Identifier 10.1109/MMUL.2021.3107862
Date of publication 27 August 2021; date of current version
6 December 2021.

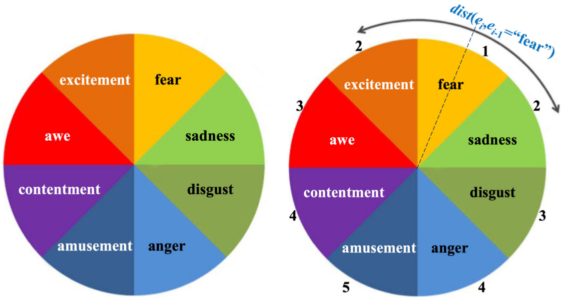

FIGURE 1. Mikels' emotion in Zhao *et al.*[4]

related studies in FER, consisting of emotion representation, well-known datasets, and typical approached. The "Emotion, Datasets, and History AFER" section details the background and practical significance of FER technology's application in different fields. The "Challenge" section discusses the challenges presented in FER's application in real scenarios and possible solutions.

## EMOTION, DATASETS, AND HISTORY

AFER is inherently a multidisciplinary enterprise involving different components, including emotion representation, datasets construction, and machine learning model. Due to the improvement of these components, FER realized a rapid development in recent years.

### Emotion Representation

We begin by briefly introducing two kinds of emotion representations frequently used in psychological research: categorical emotion states (CES) and dimensional emotion space (DES).[4] Research of emotion representation, which determines the affective information that systems are designed to detect is the basis of that of the whole FER system. CES model divides emotion into discrete categories. This model is good at matching people's experience. Thus, translating this scheme into HCI engineering framework is very convenient. The most widely applied CES model is Ekman's basic emotion categories, including anger, disgust, fear, joy, sadness, and surprise. Ekman's research is based on the cross-cultural studies, which indicates that regardless of culture differences, people perceive certain basic emotions according to FEs in the same way. Another classical model classifies emotion into eight basic categories and defines the pairwise distance among different emotion categories (see Figure 1).

However, although basic emotions can set an important criteria for FE description, there are still a large part of subtle feelings showed in daily interpersonal interactions cannot be fully demonstrated. In order to describe the emotion on a more fine-grained model, DES is proposed as an addition of CES model. DES extends discrete emotion categories to Cartesian space constituted by latent dimensions with more information.[5,6] A particular emotion can be expressed by position in 3-D or 2-D spaces like valence-arousal-dominance (VAD) and activity-temperature-weight. Unfortunately, the richness of the space makes the link of such described emotion to a FE very difficult, which is demonstrated by the related dataset size. Therefore, many FER models simplify the numerical regression in VAD to the four-class (quadrants of 2-D space) classification.

### Datasets

Recently, the extraordinary performance of machine learning methods has gradually promoted the development of FER. As the basis of machine learning, the construction of FE has also updated rapidly (see Figure 2). At the first beginning, the first FE dataset in the real sense—CK[1] is released in 2000. Although CK is limited within restricted gender, age as well as ethnic diversity, and contains only frontal views with homogeneous illumination, it is still the most important benchmark dataset that is widely used in recent 20 years. After that, different dataset captured under laboratory-controlled environment bloomed. For example, the new CK+[7] is constructed by adding spontaneous expressions and sampling numbers. Multi-PIE[2] made great progress in data diversity by including a very large number of views at different angles and diverse illumination conditions. However, traditional laboratory-controlled dataset still face two challenges. First, the laboratory environment is too monotonous. Although the model can show a good testing performance in laboratory-controlled environment, it cannot reach a satisfying result in real application scenarios, which are full of various noises. Second, the dataset size under laboratory-controlled environment cannot meet the requirements of deep learning methods, leading to the failure that the powerful feature learning ability of deep learning cannot successively applied to FER systems. To solve the aforementioned two challenges, FER dataset of hundred thousand magnitude collected from the Internet is generated. Typical FER dataset including AFEW,[8] AffectNet,[9] RAF-DB,[10] etc. Taking RAF-DB as an example, it contains not only 29,672 images, but also seven

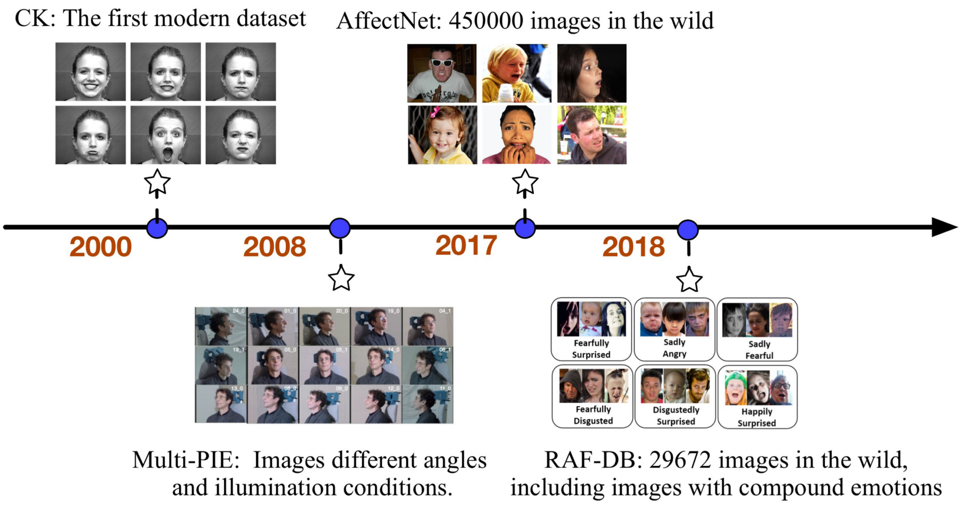
CK: The first modern dataset
AffectNet: 450000 images in the wild
2000
2008
2017
2018
Fearfully Surprised
Sadly Angry
Sadly Fearful
Fearfully Disgusted
Disgustedly Surprised
Happily Surprised
Multi-PIE: Images different angles and illumination conditions.
RAF-DB: 29672 images in the wild, including images with compound emotions

**FIGURE 2.** Historical evolution of FER datasets.

basic expressions and 12 compound expressions as labels. Thus, FER methods based on deep learning have developed a lot and gradually become the mainstream methods.

## History of Machine Learning Model for FER

As mentioned, machine learning model is the core motor of FER systems development. And the key points for FER to achieve a good performance are to extract discriminative features of different categories. In early research, the predesigned feature generated by hand-crafted methods to extract relevant information was popular. Predesigned features are usually composed of two types: appearance and geometrical. Geometrical features are mainly composed of distances, deformations, curvatures, and other geometric properties. Corresponding to this, the appearance feature uses the intensity information of the image to extract features for subsequent classification. Although this type of feature extraction method has a strong interpretability, but the poor performance still hinders its further development.

Thanks to the powerful end-to-end feature extraction, deep learning methods have risen rapidly in recent years (see Table 1). This type of method jointly optimizes feature extraction and classification weights to obtain discriminative learned feature. Deep learning methods can be split into static and dynamic, with static methods recognizing a single frame or image and dynamic ones including temporal information. Let us start with the introduction of static methods. Divided by network types, static methods can be divided into single CNN based network, multitask CNN based network, and generative adversarial networks. Single CNN based network mainly tries to enhance expression-related discriminative power of learned feature from two aspects. One is to specially design some auxiliary blocks or layers. For example, three different-level supervised blocks were proposed in Hu *et al.*[13] to obtain different levels of expression-related features. All these features are concatenated directly in scoring connection layer with the following fully connected layer. The second method is to improve the traditional softmax lost so as to increase the interclasses distance and reduce the intraclass variation of each expression class. For example, a

**TABLE 1.** Summary of representative methods for FE recognition.

| Method | Network Type | Characteristic |
|---|---|---|
| Cascade EF-GAN[11] | GAN | Better preserve identity-related features and details around eyes, noses, and mouths |
| Self-Cure Network[12] | CNN | Suppresses the uncertainties efficiently and prevents deep networks from overfitting |
| Supervised Scoring Ensemble[13] | CNN | Learns supervised scoring ensemble for FER in the wild |
| Geometry Guided Pose-Invariant GAN[14] | GAN | Allows to simultaneous facial image synthesis and pose-invariant FER |

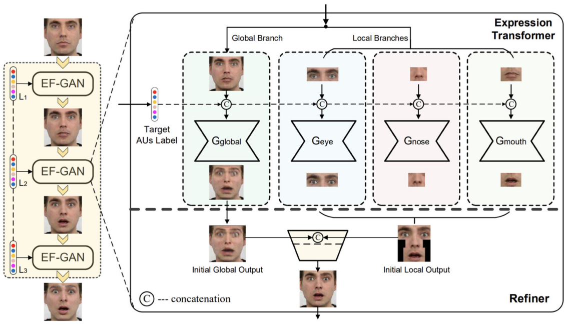

**FIGURE 3.** Upgrading the basic GAN structure[11] to generate a sufficient number of samples.

novel Logit-Weighted Cross-Entropy loss[12] is proposed to learn robust facial expression features with uncertainties. On the basis of single CNN based network, multitask CNN suggested that transferring knowledge from relevant task such as facial landmark localization is helpful for enhancing the robustness of the model. For example, geometrical features of facial landmark points is extracted to guide the performance of facial image synthesis.[14] In addition, this type of method often combines related tasks such as head pose estimation, facial action units detection, and face verification.

Apart from CNN based methods, the method of GAN has also received more and more attention in the research of recent years. The application of GAN in FER is mainly focused on generating a sufficient number of samples of different head angles, lighting environments, and expression strengths, so that subsequent deep learning models are no longer limited by the restricted dataset. A basic GAN model includes a generator used in image synthesis and a discriminator used in the quality estimation of the generated image. Two components are optimized together in the process of confronting each other until a dynamic balance is reached. For example, a large number of progressive FE editing with local expression focuses[11] can be obtained by upgrading the basic GAN structure (see Figure 3).

## APPLICATION

Current human–computer interaction (HCI) is very good at delivering specific command information, but often ignores the state analysis of implicit information about the user, especially emotional state. Such interactions are frequently perceived as cold, incompetent, and socially inept. In order to get a better interactive experience, FER is applied to all aspects of society. Next, we will introduce the application of FER in transportation, medical treatment, and education one by one.

### Transportation

With the expansion of the cities' scale and the increasing frequency of intercity exchanges, we need to use various means of transportation to shuttle between destinations every day. How to effectively prevent traffic accidents and reduce traffic risks has become an urgent problem. Dangerous driving behavior caused by extreme emotions is one of the most important causes of traffic accident. A report by the American Automobile Association Traffic Safety Foundation in July last year showed that 80% of American drivers have had "road rage" or similar behaviors in the past year, and eight million American drivers have had severe "road rage" behaviors, including dangerous behavior against other vehicles or conflict behavior against other drivers. The Foundation said that the average number of "road rage" cases that causes casualties in the United States is no less than 1200 each year. Poor driving habits, poor traffic conditions, and stress in daily life can all intensify small disputes into dangerous "road rage." Taking the case of May 2015 as an example, a car driver knocked a motorcycle driver in to the air at a speed of 130 km/h due to unpleasant driving, and eventually caused the latter's death due to ineffective rescue. In response to the above problem, many companies have applied FER technology to develop emotion monitoring systems in the cockpit. Take Affectiva as an example. The system developed by the company can recognize the emotions of people in the car in real time, and offer alert when the driver is too tired.

### Medical Treatment

According to WTO statistics on the global burden of disease in 2017, depression will become the second largest source of disease burden after cardiovascular diseases. Correspondingly, the 2019 China Mental Health Survey Research Report shows that the prevalence of depression in China has reached 0.068. However, the clinical diagnosis of similar mental diseases almost relies on clinical interviews and questionnaires, lacking long-term quantitative indicators, which has led to a large number of misdiagnosis and missed diagnosis. Recently, the rapid development of automatic FER is expected to solve these problems. A significant correlation between features such as movement changes of eyes, eyebrows as well as corners of mouth and depression is found through the technical analysis of FER.[15] In addition, many medical systems have been able to effectively mine and encode facial emotion information, so as to make a more accurate assessment of the degree of depression.

### Business

FER FE technology provides the possibility to automatically analyze the mental state of users. Being able to correctly understand the user's preferences for the product is of great significance and value for subsequent business decisions. Many related companies have created sentiment analysis systems for a majority of filmmakers and advertising producers. By analyzing the impact of movies or advertisements on consumers, a clear direction is proposed for product improvement, thereby improving the entire narrative process. In addition to perceiving the emotions of consumers, some companies provide indispensable help for editing and viewing guidance by identifying the emotions of characters in multimedia.

## CHALLENGE

### Class Incremental

With the rapid expansion of social network data and the continuous refinement of emotional research in psychology, more incremental and fine-grained expression classes are being discovered and defined nowadays. Compared to the traditional six-class FE system, these incremental expression types can describe facial muscle movements in more detail and show facial emotions more realistically. In some special environments such as medical and psychological research, it is inevitable as well that expression types will gradually increase. In such scenarios, how to deal with the incremental classes becomes the key point. This problem is easy to solve by retraining the model from scratch if large expression dataset, infinite computing resources, and adequate storage are accessible. But in real applications, these conditions are almost impossible to meet. Additionally in certain applications, due to restrictions of copyright and privacy policy, it is usually difficult to access all the data of the expression types that have participated in the training during the process of model updating. Therefore, class incremental learning methods attract a lot of attention in the field of FE recognition. Catastrophic forgetting is considered to be a major issue in incremental learning, as most learning methods are tended to gradually lose the memory of old expression classes when new expression data arrives in the model iteration steps. In traditional machine learning research, some classic methods were proposed to overcome catastrophic forgetting problem and are currently applied in many applications. In incremental SVM learning, Kuhn–Tucker (KT) conditions are modified to keep seen classes in the training iteration of one vector each time. Nearest class mean (NCM) method is another famous method to learn with a fixed data representation. In NCM classifier, a prototype vector is calculated to represent each class and keeps being updated incrementally through the dataflows. NCM has been used frequently in metric learning and performs well in certain tasks such as large scale image classification. For deep learning-based methods, there are three ideas to overcome the catastrophic forgetting problem. The first idea focuses on modifying the network structure and parameter settings of the deep model. Growing a brain increases model capacity by extending existing layers or adding new layers to allow for more natural model adaptation. Deep adaptation modules (DAMs) is proposed to augment a network by incorporating filters incrementally and limiting those newly learned filters to linear combinations of old ones. The second idea is to retain the design structure of network parameters unchanged and propose various methods to avoid catastrophic forgetting in the procedure of fine-tuning previous deep models. Learning without forgetting (LwF) exploits to incorporate standard cross-entropy loss with the distillation loss to preserve existing knowledge using only new train data in the process of model iteration. Inspired by LwF method, iCaRL builds a bounded exemplar memory based on a combined loss function of classification and distillation losses, and proposes a nearest-exemplar-mean (NEM) classifier to handle the class imbalance problem. The third idea explores to learn fix representations as feature extractor at the beginning of model iteration. Deep-shallow incremental learning learns an initial fix representation and trains independent shallow classifiers continually as new classes come.

### Open Set Recognition (OSR)

In real-world expression recognition scenarios, unseen classes are quite common in model testing because the number of expression categories is always in constant refinement and dynamic growth. In some tasks, only certain expressions are necessary to be classified correctly while the other expressions are not paid attention to. On these conditions, it is usually difficult as well as costly to exhaust all expression classes during training. A more realistic solution is to admit that quite a few classes cannot be collected during training process and propose

OSR methods that are able to deal with those unseen classes robustly and effectively. Two basic categories of classes, known known classes (KKCs) and unknown unknown classes (UUCs), should be considered in the OSR scenario. KKCs are the classes with precisely labeled training samples. UUCs are the classes without any initial information during training. To reject UUCs existing in open world is a major issue in the OSR task, which requires strong generalization of the recognition model. Discriminative model and generative model are the mostly applied techniques to solve this problem. From the discriminative perspective, both traditional machine learning-based and deep learning-based models are proposed to adapt original methods to the OSR setting. 1-vs-Set Machine are introduced as a new SVM to address overgeneralization risk and overspecialization risk and form a decision space based on the marginal distances of a normal SVM for every KKCs. Sparse representation-based open set recognition (SROSR) uses the extreme value theory (EVT) to model the hidden discriminative information in the tail part of the matched reconstruction errors and the sum of nonmatched reconstruction errors to turn the OSR problem into a set of statistical hypothesis testing problems. Nearest non-outlier (NNO) adapts the nearest class mean (NCM) classifier using a well-designed measurable recognition function for open world recognition to achieve a balance of open space risk and classification accuracy. Extreme value machine (EVM) is derived from statistical extreme value theory (EVT) and takes a great advantage over other classifiers in the deep feature space. OpenMax[16] first explores the OSR solution in deep networks by replacing the standard SoftMax function with its novel OpenMax layer to output the estimated probability of KKCs and UUCs, respectively. From the generative perspective, both instance and noninstance methods are proposed. Generative OpenMax (G-OpenMax)[17] extends OpenMax with generative adversarial networks to synthesis samples of unknown classes and provides brief visualization of unknown classes in open space. Collective decision-based open set recognition (CD-OSR) introduces a novel collective/batch decision strategy and modifies hierarchical Dirichlet process to offer a solution for the OSR problem of collective decision without setting decision threshold.

## Open Set Recognition

Faces in the open world are sometimes partially occluded in some real scenarios. Occluded expression recognition is an inherent challenge and a hot spot for facial analysis tasks. Facial occlusion usually leads to poor performance for existing expression classifiers and brings up a significant obstacle to open-world FE recognition task.

Knowledge from several related topics has been explored to solve the occlusion problem. A deep-structure algorithm is proposed to address four frequently occurred occlusion conditions.[18] This method uses Gabor filter as its feature extractor and a multilayers network as the pretrain model. Transfer learning is used to tackle the severe occlusion of the upper half of the face in a VR setting [15]. Deep belief network is utilized as a generative model to capture pixel-level features with precise indications of captured information in each layer. By learning a deep generative model, this method is able to fill in occluded regions so that original expressions can be reconstructed.[19] Convolutional neural network with attention mechanism (ACNN) helps the model to focus on the most discriminative unoccluded areas instead of getting lost in that meaningless information. Detection of occlusions is not explicitly needed in an ACNN model, preventing propagating any detection mistakes afterward. Sparse representation is incorporated to form a general recognition algorithm, which can handle occlusion and corruption errors uniformly and provide an estimate of the limit size of acceptable occluded areas based on the theory of sparse representation.[20] Mutually reweighted $L_1$ regularization term is applied to the training process on unoccluded expression samples and ensures that the resulting classifier is highly robust to occlusions.

## ACKNOWLEDGMENTS

This work was supported in part by National Key Research and Development Project under Grant 2018YFB1703404, and in part by the NSFC Program under Grants U1801263,62076146, U20A600004, U19A2062, and U1911401.

## REFERENCES

1. T. Kanade, J. F. Cohn, and Y. Tian, “Comprehensive database for facial expression analysis,” in *Proc. 4th IEEE Int. Conf. Autom. Face Gesture Recognit. (Cat. No. PR00580)*, 2000, pp. 46–53.
2. R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, “Multi-pie,” *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, 2010.
3. S. Li and W. Deng, “Reliable crowdsourcing and deep locality-preserving learning for unconstrained facial expression recognition,” *IEEE Trans. Image Process.*, vol. 28, no. 1, pp. 356–370, Jan. 2019.

4. S. Zhao *et al.*, "Predicting personalized emotion perceptions of social images," in *Proc. 24th ACM Int. Conf. Multimedia*, 2016, pp. 1385–1394.
5. S. Zhao *et al.*, "Affective image content analysis: Two decades review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, to be published, doi: 10.1109/TPAMI.2021.3094362.
6. S. Zhao *et al.*, "Discrete probability distribution prediction of image emotions with shared sparse learning," *IEEE Trans. Affective Comput.*, vol. 11, no. 4, pp. 574–587, Oct.–Dec. 2018.
7. 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 *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.-Workshops*, 2010, pp. 94–101.
8. A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE Multimedia*, vol. 19, no. 3, pp. 34–41, Jul.–Sep. 2012.
9. A. Mollahosseini, B. Hasani, and M. H. Mahoor, "AffectNet: A database for facial expression, valence, and arousal computing in the wild," *IEEE Trans. Affective Comput.*, vol. 10, no. 1, pp. 18–31, Jan.–Mar. 2017.
10. S. Li, W. Deng, and J. Du, "Reliable crowdsourcing and deep locality-preserving learning for expression recognition in the wild," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2017, pp. 2584–2593.
11. R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 5021–5030.
12. K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2020, pp. 6897–6906.
13. P. Hu, D. Cai, S. Wang, A. Yao, and Y. Chen, "Learning supervised scoring ensemble for emotion recognition in the wild," in *Proc. 19th ACM Int. Conf. Multimodal Interact.*, 2017, pp. 553–560.
14. F. Zhang, T. Zhang, Q. Mao, and C. Xu, "Geometry guided pose-invariant facial expression recognition," *IEEE Trans. Image Process.*, vol. 29, pp. 4445–4460, Feb. 2020.
15. Q. Wang, H. Yang, and Y. Yu, "Facial expression video analysis for depression detection in Chinese patients," *J. Vis. Commun. Image Representation*, vol. 57, pp. 228–233, 2018.
16. A. Bendale and T. E. Boult, "Towards open set deep networks," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, 2016, pp. 1563–1572.
17. Z. Ge, S. Demyanov, Z. Chen, and R. Garnavi, "Generative openmax for multi-class open set classification," in *Proc. British Mach. Vis. Conf.*, 2017.
18. Y. Cheng, B. Jiang, and K. Jia, "A deep structure for facial expression recognition under partial occlusion," in *Proc. 10th Int. Conf. Intell. Inf. Hiding Multimedia Signal Process.*, 2014, pp. 211–214.
19. M. Ranzato, J. Susskind, V. Mnih, and G. Hinton, "On deep generative models with applications to recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2011, pp. 2857–2864.
20. J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

**XIBIN ZHAO** is an Associate Professor with the School of Software, Tsinghua University, Beijing, China. His research interests include reliability analysis of affective computing and information system security. He received the B.S., M.E., and Ph.D. degrees in computer science from the School of Computer Science and Telecommunication Engineering, Jiangsu University, Zhenjiang, China, in 1994, 2000, and 2004, respectively. Contact him at zxb@tsinghua.edu.cn.

**JUNJIE ZHU** is currently working toward the Ph.D. degree. His main research interests include machine learning and affective computing. Contact him at zhujj18@mails.tsinghua.edu.cn.

**BINGJUN LUO** is currently working toward the Ph.D. degree. His main research interests include machine learning and affective computing. Contact him at luobingjun@gmail.com.

**YUE GAO** is an Associate Professor with the School of Software, Tsinghua University. He has also been working with the School of Computing, National University of Singapore, Singapore, and the Medicine School, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA. He received the B.S. degree from the Harbin Institute of Technology, Harbin, China, and the M.E. and Ph.D. degrees from Tsinghua University, Beijing, China. Contact him at gaoyue@tsinghua.edu.cn.