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.2022.3143100, IEEE Transactions on Affective Computing

An Overview of Facial Micro-Expression Analysis: Data, Methodology and Challenge

Hong-Xia Xie, Ling Lo, Hong-Han Shuai, and Wen-Huang Cheng, *Senior Member, IEEE*

Abstract—Facial micro-expressions indicate brief and subtle facial movements that appear during emotional communication. In comparison to macro-expressions, micro-expressions are more challenging to be analyzed due to the short span of time and the fine-grained changes. In recent years, micro-expression recognition (MER) has drawn much attention because it can benefit a wide range of applications, e.g. police interrogation, clinical diagnosis, depression analysis, and business negotiation. In this survey, we offer a fresh overview to discuss new research directions and challenges these days for MER tasks. For example, we review MER approaches from three novel aspects: macro-to-micro adaptation, recognition based on key apex frames, and recognition based on facial action units. Moreover, to mitigate the problem of limited and biased ME data, synthetic data generation is surveyed for the diversity enrichment of micro-expression data. Since micro-expression spotting can boost micro-expression analysis, the *state-of-the-art* spotting works are also introduced in this paper. At last, we discuss the challenges in MER research and provide potential solutions as well as possible directions for further investigation.

Index Terms—Facial Micro-expression, Recognition, Spotting, Action Units, Deep Learning, Survey.

1 Introduction

Facial micro-expression (ME) is a result of *conscious suppression* (intentional) or *unconscious repression* (unintentional), which can be viewed as a "leakage" of people's true feelings [1]. MEs are brief involuntary facial expressions that usually appear when people are trying to conceal their true feelings, especially in high-stake situations. Hence, micro-expression recognition (MER) research enables greater awareness and sensitivity to subtle facial behaviors, and is an important subject for human emotion and affective phenomena understanding, which has been explored by various disciplines such as psychology, sociology, neuroscience, computer vision, etc. Such skills are useful for psychotherapists, interviewers, and anyone working in communications.

1.1 The Difference between Macro-expression (MaE) and Micro-expression (ME)

Micro-expressions occur when people are trying to conceal or repress their true feelings [5], [6]. On the contrary, macro-expressions are the normal facial expression of emotions and are easy to be perceived and interpreted by others in daily interactions. The major difference between macro-expressions and micro-expressions lies in their intensity and duration. Although there is no strict rule of the threshold to distinguish one from the other, most agreed that macro-expressions usually last from 0.5 to 4 seconds while micro-expressions should be no longer than 0.5 seconds [7]. Most researches [8], [9], [10] set 0.5 seconds as the threshold, while 0.2 seconds duration is also regarded as a boundary for differentiating micro- and macro-expressions as validated in [11]. Besides, the neural mechanisms underlying the recognition of micro-expression and macro-expression are different, showing different electroencephalogram (EEG) and event-related potentials (ERPs) characteristics. The brain regions responsible for their differences might be the inferior temporal gyrus and the frontal lobe [11]. Generally,

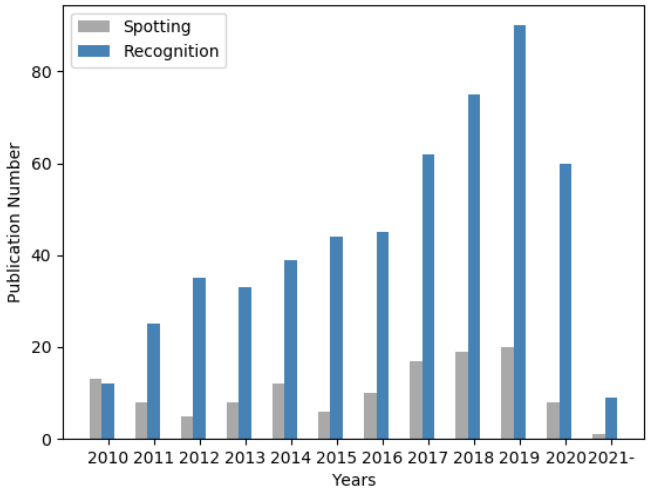

Fig. 1. The publication number of Micro-expression Recognition and Spotting from 2010 to May 2021 (Source: Web of Science).

- *H.-X. Xie, L. Lo are with the Institute of of Electronics, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan. E-mai: {hongxiaxie.ee08g, lynn97.ee08g}@nctu.edu.tw.*
- *H.-H. Shuai is with the Department of Electrical and Computer Engineering, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan. E-mail: hhshuai@nctu.edu.tw.*
- *W.-H. Cheng is with the Institute of Electronics, National Yang Ming Chiao Tung Univresity, Hsinchu, 300 Taiwan, and the Artificial Intelligence and Data Science Program, National Chung Hsing University, Taichung, 400 Taiwan. E-mail: whcheng@nctu.edu.tw.*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

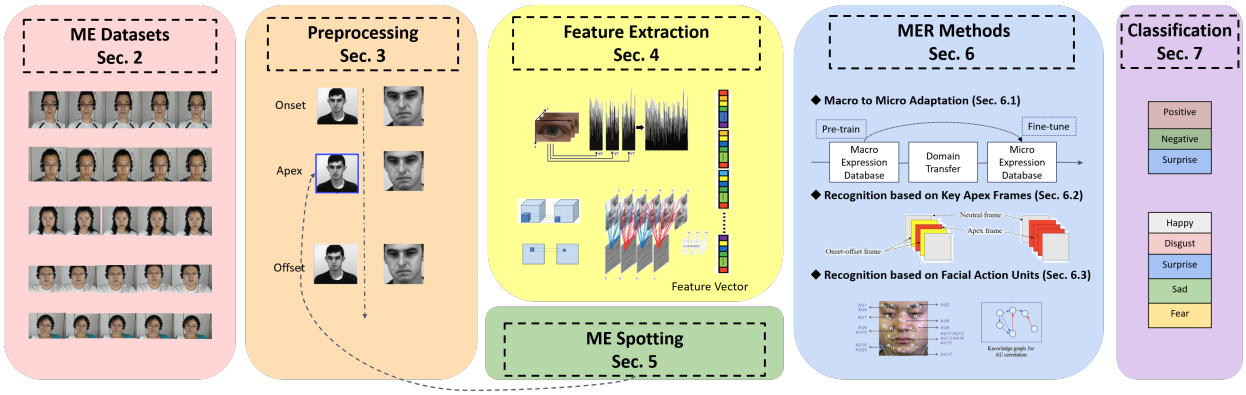

Fig. 2. The organization of this survey is structured according to the general MER pipeline.

TABLE 1
Academic challenges for micro-expression recognition and spotting.

Challenge	Year	Dataset	Task	Evaluation Metric	Event
MEGC2018 [2]	2018	CASME II, SAMM	Cross-database (HDE, CDE)	UAR, WAR, F1-score	FG2018[1]
MEGC2019 [3]	2019	CAS(ME)^2, SAMM	Recognition, Spotting (CDE)	UAR, WAR, F1-score	FG2019[1]
MEGC2020 [4]	2020	CAS(ME)^2, SAMM Long Video	Spotting	F1-score	FG2020[1]
MER2020*	2020	Synthetic Data	Recognition	F1-score	ICIP2020[2]
MEGC2021	2021	CASME II, SAMM, SMIC	Data Generation, Spotting	Experts	ACM MM[3]

* MER2020: http://mer2020.tech/
[1] FG: IEEE International Conference on Automatic Face and Gesture Recognition
[2] ICIP: IEEE International Conference on Image Processing
[3] ACM MM: ACM Multimedia (https://megc2021.github.io)

the macro-expression shows higher intensity and visibility when compared to the micro-expression.

Different from recognizing the macro-expressions, the short duration and low intensity of micro-expression create more challenges and raise in the level of difficulty and to recognize them. Unlike interpreting macro-expressions, the pre-processing stage becomes vital to deal with these additional challenges. For example, the subtle muscle movements of micro-expression are magnified for more accurate recognition. Moreover, in order to identify the brief occurrence of micro-expressions in long videos, micro-expression spotting approaches have also been explored for subtle muscle movements locating as well as filtering the redundant information in the raw videos.

1.2 Facial Action Coding System

Facial Action Coding System (FACS) was first proposed by Ekman and Friesen and updated in 2002 [12], [13] to factorize the composition of micro-expressions. It is the most widely used coding scheme for decomposing facial expressions into individual muscle movements, called Action Units (AUs). With FACS, every possible facial expression can be described as a combination of AUs. There are 32 facial muscle-related actions, and 6 extra unspecific miscellaneous Action Descriptors (ADs) [14].

Yet, the professional training of FACS encoding experts is time-consuming. Professional encoders usually need to receive 100 hours of training, and in practice the encoding process takes 2 hours to encode a 1-minute video on average [15]. Thus, an automatic AU recognition system with high accuracy can be very helpful and valuable.

Since AUs are descriptive for certain facial configurations, specific systems are proposed to explore the relationship between facial muscle movements (AUs) and human emotions, e.g., Emotional Facial Action Coding System (EMFACS-7) [16], Facial Action Coding System Affect Interpretation Dictionary (FACSAID) [17] and the System for Identifying Affect Expressions by Holistic Judgments (Affex) [18]. It can be found that various mapping strategies for AUs and emotions are adopted by existing facial expression datasets due to the lack of standard guidelines [19].

1.3 Academic Challenges for MER

An Academic challenge is a competition created by academic experts for leveraging the power of open innovation to advance the *state-of-the-art* in a particular field. Within the computer vision domain, several academic challenges related to micro-expression recognition and spotting have been proposed. Some known events in recent years are summarized in Table 1.

1.4 Outline of this Paper

Numerous works have been accomplished to mitigate the challenging MER task. The publication counting from 2010 to May 2021 is shown in Fig. 1. Several surveys on MER have been published in recent years [24], [25], [26], [27], [28], [29], [30], [23], [31]. However, most of them have been focused on traditional image-processing methods and a fresh overview

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 2
A list of public datasets on spontaneous micro-expression.

Dataset		Macro/Micro	Videos	FPS	Resolution	FACS	Emotion	Subjects	AU	Index*	Ethnicities	Mean Age
CASME [9]		micro	195	60	640×480,1280×720	Yes	8	19	No	On,Apex,Off	1	22.03
SMIC [20]	HS	micro	164	100	640×480	No	3	16	No	On,Off	3	26.7
	VIS		71	25				8		N/A		
	NIR		71	25				8		N/A		
CASME II [10]		micro	247	200	640×480	Yes	5	26	Yes	On,Apex,Off	1	22.03
SAMM [8]		micro	159	200	2040×1088	Yes	7	32	Yes	On,Apex,Off	13	33.24
CAS(ME)^2 [21]		macro & micro	300 & 57	30	640×480	No	4	22	N/A	On,Apex,Off	1	22.03
SAMM Long Videos [22]		macro & micro	343 & 159	200	2048×1088	Yes	N/A	30	Yes	On,Apex,Off	13	33.24
MMEW [23]		macro & micro	900 & 300	90	1920×1080	Yes	7	36	Yes	On,Apex,Off	1	22.35

* On, Apex, Off: Onset frame, Apex frame, Offset frame, respectively.

to discuss new directions and challenges the MER faces today is necessary. In this survey paper, therefore, we provide a more comprehensive and in-depth study of existing approaches, e.g. including over 100 papers for MER until May 2021 which have not been reviewed in the previous survey papers. Based on the nature of MEs, e.g., the short span of time and the fine-grained changes, we introduce MER approaches from three novel aspects: macro-to-micro adaptation, recognition based on key apex frames, and recognition based on facial action units. Moreover, to mitigate the problem of limited and biased ME data, synthetic data generation algorithms are surveyed for the diversity enrichment of ME data. To the best of our knowledge, this is the first review to summarize the synthetic data generation solutions based on spontaneous ME datasets. Since micro-expression spotting can boost micro-expression analysis, the *state-of-the-art* spotting works are also introduced. At last, we discuss open challenges and provide potential solutions as well as future research directions.

The whole systematic framework of MER is present in Fig. 2. The outline of this paper adheres as follows: Section 2 introduces ME data collection, current public ME datasets, and popular data synthesis methods. Section 3 describes pre-processing techniques commonly used for ME data. Section 4 provides a detailed review of feature extraction methods based on handcrafted features and deep learning-based methods. Discussions of ME spotting are present in Section 5. Section 6 summarizes the commonly-used techniques on MER recently. The loss function widely used for MER is introduced in Section 7. Section 8 presents the overall performance comparison. Open MER challenges and the possible solutions are arranged in Section 9 and some potential future directions for research are suggested in Section 10.

2 DATASETS

2.1 ME Data Collection

Emotion Classes. In discrete emotion theory, there are many different basic emotion definition systems, of which the most popular one adopted in the computer vision community is conducted by Ekman [32], where the basic emotions are divided into six categories of anger, disgust, fear, happiness, sadness, and surprise. An extra emotion of contempt is added in later researches.

Current public ME datasets divide micro-expression emotions into different classes according to their collection strategy. The emotion labels should take into account AUs, participants' self-report as well as the stimuli video content, etc. Class number of different ME datasets are provided in Table 2. Fig. 3 shows sample frames from current ME datasets.

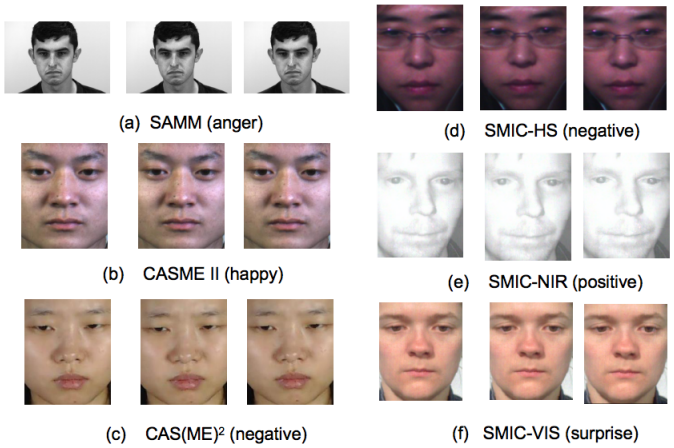

Fig. 3. Sample frames along with emotion labels from different ME datasets.

2.1.1 Apparatus Setup

Existing ME datasets are usually collected in a strictly lab-controlled environment with different frame rates and resolution setting. The observers are kept out of sight to maximise the chances of natural suppression by making participants as comfortable as possible.

2.1.2 Micro-expression Inducement

Inducing micro-expressions effectively is an tricky step during data collection.

Emotional stimuli. MEs are more likely to occur under high-arousal stimuli, so video clips with high emotional valence are proved to be effective materials for eliciting MEs [8], [9], [10], [20]. While recently proposed MMEW [23] adopted another strategy by constructing a high-stakes situation, i.e., poker games or TV interviews with difficult questions.

Emotion Eliciting. Based on Ekman and Friesen [6], there are two types of micro displays that may reveal intrinsic feelings: "The time-reduced full affect micro displays (i.e., micro-expressions) may well be those which the ego is not aware of, while the squelched micro displays may

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

be those which the ego senses and interrupts in mid performance." A ME dataset collected based on the above theory is called **spontaneous**. Accordingly, current spontaneous ME datasets collect data following two elicitation paradigms: asking participants to fully suppress facial movements or suppress facial movements when they are self-aware. Specifically, SMIC, CASME, and SAMM asked participants to fully suppress facial movements during the whole experiment so that MEs may occur. For CASME II, half of the participants were asked to keep neutral faces when watching video clips while other participants were enforced to suppress the facial movements when they realize there is a facial expression. Notably, participants were needed to keep neutral faces since watching emotional video episodes with neutralizing faces is effective to elicit spontaneous MEs without many irrelevant facial movements [9].

By contrast, **non-spontaneous** databases, for example, Polikovsky's Database [33] and USF-HD [34] data were collected by asking participants to pose or mimic facial expressions which contradict the spontaneous nature of ME. Specifically, in USF-HD, participants were asked to perform two types of micro-expressions, while in Polikovsky's Database, participants were asked to simulate micro-expression motions.

Emotion Label Rating Quantification. To explicitly quantify the personal emotional response to the emotional stimulus, a predefined scale that a subject can assign to the perceived emotion response is usually needed. For example, SMIC [20], CASME [9], and CASME II [10] databases assigned an emotion label to stimuli videos based on self-reports completed by participants.

Not All Subjects Show Micro-expressions. According to Ekman's research, when people are telling lies, about half of them might show MEs, while the other half do not [5]. The reason why only some people show MEs is still unclear. In SMIC, four participants did not show any ME at all throughout the course of 35-minutes video watching.

2.2 Spontaneous Micro-Expression Dataset

Since posed ME datasets are collected by intentionally controlled, it contradicts the natural occurrence of MEs. In this paper, we only focus on **spontaneous ME datasets**. Table 2 provides an overview of spontaneous ME datasets.

Chinese Academy of Sciences Micro-Expression (CASME) [9] dataset consists of 195 video samples from 19 valid subjects with a frame rate of 60 fps. The samples for 8 emotions are highly imbalanced (5 happiness, 6 sadness, 88 disgust, 20 surprise, 3 contempt, 2 fear, 40 repression and 28 tense). Notably, in CASME, each sample is recorded with two different cameras and environmental settings, namely Class A and Class B. Class A was recorded by BenQ M31 camera with natural light, while class B was by GRAS-03K2C camera and two LED lights. All participants are Chinese.

Chinese Academy of Sciences Micro-Expression II (CASME II) [10] is an improved version of CASME collected in a well-controlled lab environment. It contains 247 ME sequences from 26 subjects with 7 categories, including happiness, disgust, surprise, fear, sadness, repression and others, which were labeled based on AUs, participants' self-report and the content of stimuli videos. All subjects are Chinese. The participants are limited to youths from China.

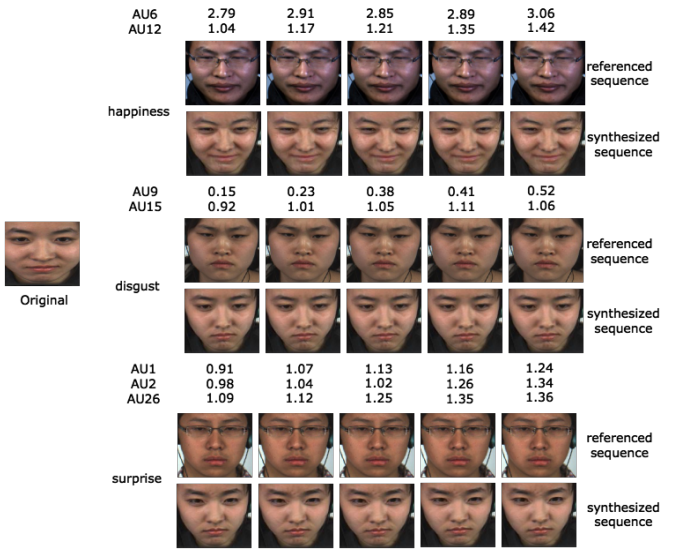

Fig. 4. Synthetic faces generated by AU-ICGAN from CASME II dataset. The number above each image indicates the target AU intensities of chosen AUs [35].

Spontaneous Micro-expression Corpus (SMIC) [20] is composed of 164 ME sequences from 16 subjects filmed at 100 fps. It is one of the first to include spontaneous MEs through emotional inducement experiments. For data reliability, ME clips in SMIC are classified into three classes: positive, negative and surprise. While positive (happy) and surprise only include one target emotion each, negative includes four emotions (sad, anger, fear and disgust). The subjects distribution is ten Asians, nine Caucasians and one African participant, with mean age of 26.7.

Spontaneous Actions and Micro-Movements (SAMM) [8] has 159 ME sequences from 32 subjects. The seven classes include contempt, disgust, fear, anger, sadness, happiness and surprise. To enlarge the ethnic diversity, participants recruited by SAMM have 13 ethnicities: 17 White British, three Chinese, two Arab, two Malay and one each: African, Afro-Caribbean, Black British, White British/Arab, Indian, Nepalese, Pakistani and Spanish. An even gender split was also achieved, with 16 male and 16 female participants. The ethnicities of participants are diverse and the gender split is even. Moreover, SAMM has advantage over others in age distribution with mean age of 33.24 (SD: $\pm$ 11.32), while CASME and CASME II have a mean age of 22.03 (SD: 1.60).

It should be noted that most of the existing datasets usually obtain subjects from few countries and contain a majority bias on ethnicities. In addition, the standard deviations of the age distribution are usually low. For subjects from different culture background or in different age group, their facial expressions may differ from each other. Though the differences might be slight, the subtle changes are vital for micro-expression recognition. The bias of age and ethnicity lie in the dataset tend to prevent the recognition methods from finding the general solution and the performance would be affected. Especially on the cross-database evaluation, the solution for a dataset lacking of

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

diversity would be difficult to generalize to other datasets or even real-world scenario.

2.3 Macro- and Micro-Expression Datasets

Considering real-world scenarios, macro- and micro-expressions could co-occur, there are three public datasets combining macro and micro facial expressions, i.e., **SAMM Long Videos [22]**, **CAS(ME)^2 [21]** and **MMEW [23]**. All of them can be empolyed for macro-expression and micro-expression spotting from long videos.

The Chinese Academy of Science Macro- and Micro-expression (CAS(ME)^2) dataset was established by the Chinese Academy of Science. The dataset contains 300 macro-expressions and 57 micro-expressions, with four different emotional labels: positive, negative, surprise and others from 22 participants (13 females and 9 males).

SAMM Long Videos consists of 147 long videos with 343 macro-expressions and 159 micro-expressions. The frame rate is 200 fps, and 0.5 seconds is set as the threshold for classifying macro- ($\geq$ 0.5 seconds) and micro-expressions ($<$0.5 seconds). Emotion labels are not provided.

Micro-and-Macro Expression Warehouse (MMEW) is the largest macro-and-micro dataset up to date, which includes 900 macro- and 300 micro-expressions. The expression samples are coded with the onset, apex, and offset frames, with AUs marked and emotions labeled in both CAS(ME)^2 and MMEW.

2.4 Synthetic Data Generation

While building ME databases, it is not only challenging to trigger an ME but also very difficult to label one. Labeling ME data requires human labor as well as professional domain knowledge. Even with professional training, it is reported that up to only 47% labeling accuracy can be achieved for a human expert [36]. Thus the size of existing databases is usually very limited. Synthetic datasets are introduced in this situation when annotating the ground-truth is a time-consuming work. Notably, synthesizing facial data is more challenging than other basic objects as the fidelity of human faces is hard to preserve [37]. Queiroz *et al.* [38] presented a methodology for generation of facial ground truth with synthetic faces. The 3D face model database can control face actions as well as illumination conditions, allowing to generate animation with different facial expressions and eye motions with the ground truth of the facial landmark points provided at each frame. Abbasnejad *et al.* [39] established a large-scale dataset of facial expression using a 3D face model. It consists of shape and texture models to create different subjects with different expressions. Their experimental results showed that the synthesized dataset enables efficiently deep network training for expression analysis.

Apart from synthetic databases generated by 3D morphable models, an alternative way to produce synthetic data is using Generative Adversarial Network (GAN). ExprGAN [40] was the first GAN-based model that can transform the face image to have a new expression where the expression intensity can be adjusted continuously. To address the generated artifacts and blurs around expression-intensive regions, Cascade EF-GAN [41] was proposed to perform progressive facial expression editing with local expression focuses. Zhang *et al.* [42] took advantage of GAN and proposed a network exploiting different poses and expressions jointly for facial image synthesis and pose-invariant facial expression recognition. The generated face images with different expressions under arbitrary poses can enlarge and enrich expression datasets and benefit the recognition accuracy. Cai *et al.* [43] synthesized images with a same face in different expressions using a conditional generative model. The resulting dataset consists of sets of images and each image set contains the same identity with different synthetic expressions, which benefits the identity-free expression recognition.

Different from large-scale facial expression, Xie *et al.* [35] proposed AU Intensity Controllable Generative Adversarial Networks (AU-ICGAN) for micro-scale expression data synthesis. To enrich the limited data samples, their AU-ICGAN aims to generate face images with specific intensities of action units to simulate real-world ME data. In addition, the image structure similarity along with image sequence authenticity is taken into consideration so that the generated ME image sequences can be more realistic. Fig. 4 shows the synthetic results for enriching ME datasets. Their experimental results showed that pretrained on an auxiliary synthetic dataset can benefit deep MER networks. However, as indicated in [44], undesirable generated large-scale synthetic data may exist dataset bias, which would harm to face recognition system. More research is needed to further explore how the quality of synthetic data affects MER system.

The synthetic ME data can be evaluated from objective and subjective views. For subjective evaluation, the Facial Micro-Expression (FME) Workshop and Challenge 2021 [1] encourages advanced techniques for facial expressions generation and spotting, where each generated image will be evaluated based on the quality and action units by three experts. However, this kind of process may be time-consuming. For objective evaluation, to ensure the generated synthetic data has a high quality to replace or supplement the real data, it is useful to train learning-based models on generated data and test on real data whilst obtaining desirable performance, as in AU-ICGAN [35]. Moreover, there are fidelity metrics designed for synthetic data, including general Image Quality Assessment (IQA) and face-specific quality evaluation (Face Quality Assessment, FQA) [45]. The former contains Mean Square Error (MSE), Peak Singal-to-Noise Ratio (PSNR), Structure Similarity Index Measure (SSIM), etc. While the latter includes Inception Score (IS), Fechit Inception Distance (FID) which measures the KL-divergence distance of generated and real data, etc.

3 Pre-processing

The pre-processing stage in MER consists of all steps required before the extraction of meaningful features can commence. One of the important aim of the pre-processing stage is to detect and align faces into a common reference, so that the features extracted from each face correspond to the same semantic locations. It removes rigid head motion and, to some extent, the anthropomorphic variations among people.

1. https://megc2021.github.io

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

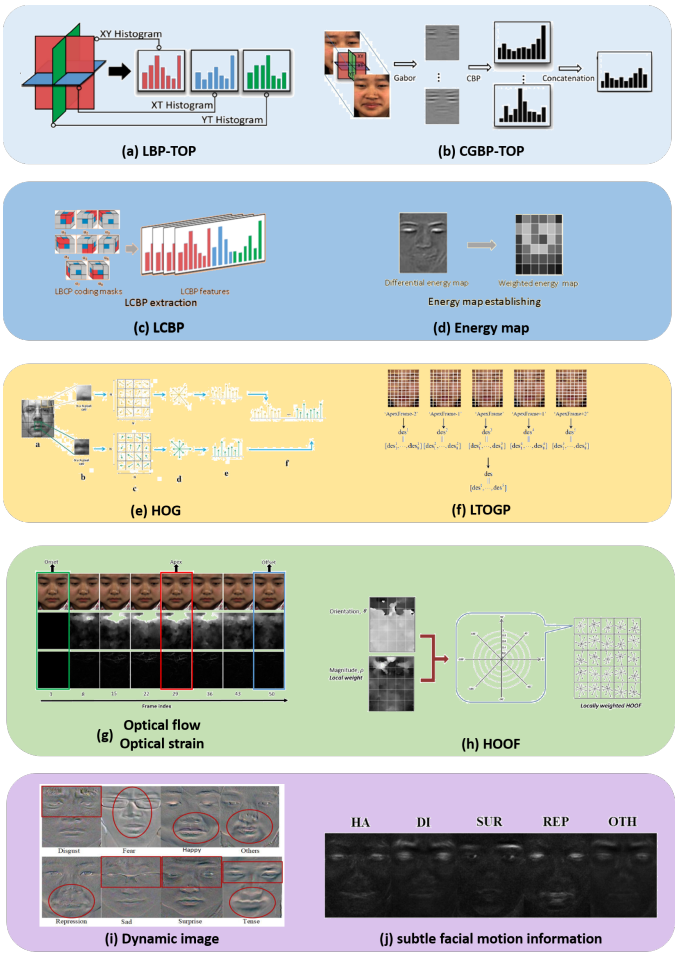

Fig. 5. An illustration of handcrafted features: (a) LBP-TOP [46], (b) CGBP-TOP [47], (c) LCBP [46], (d) energy map [46], (e) HOG [48], (f) LTOGP [49], (g) optical flow (optical strain) [50], (h) HOOF [50], (i) dynamic image [51] and (j) subtle facial motion information [52].

After alignment, the subtle micro muscle movements can be further magnified in order to enhance the discriminative characteristic. Please refer to our supplementary materials for more details.

4 Feature Extraction

Feature representation plays an important role for MER. With proper extraction, the raw input data of a ME video clip can be represented in a simple and concise form. The two main challenges for the feature descriptor of ME are 1) It should be able to capture the difference in both spatial and temporal domains and 2) It should be able to capture the micro-level differences. In our survey, we divide the commonly used feature representations into two main categories: handcrafted feature and learning-based feature.

4.1 Handcrafted Feature

Handcrafted feature for face data can be further divided into appearance-based feature and geometry-based feature [61]. Appearance-based feature represents the intensity or the texture information of face region data while geometry-based feature describes the face geometrics such as the location of each facial landmark. In the literature, appearance-based feature has shown its effectiveness on dealing with illumination difference and unaligned images [62], [63], [64], [65]. An illustration of handcrafted features is shown in Fig 5.

Local Binary Pattern (LBP)-based Feature. LBP-based feature is commonly used in the literature due to the computational simplicity. The LBP was first proposed in [66] as a texture operator via thresholding the eight neighbors of each pixel and representing the result with a binary code. An extended version of LBP is described in [67] to meet the need of rotation invariance.

However, for an ME video clip, the dynamic information between different frames is crucial. As a time-domain extension of LBP, LBP-TOP was proposed to deal with dynamic feature analysis [68]. Usually, LBP-TOP features are applied with region-based algorithms to improve the robustness of misalignment. Currently, LBP-TOP is reported by most of the existing ME datasets as the baseline evaluation. Due to its computational simplicity, LBP-TOP features are utilized in a variety of different MER frameworks including cross-domain MER and ME spotting [69], [70], [56].

LBP-TOP variants. Aside from the use of original LBP-TOP operator, several variants were proposed to meet different needs for MER [52], [78], [79], [80], [47], [46], [81]. Huang *et al.* [52] combined the idea of integral projection and texture descriptor like LBP-TOP to bone texture characterization and face recognition. Guo *et al.* [79] extended the LBP-TOP operator into two novel binary descriptors: Angular Difference LBP-TOP (ADLBPTOP) and Radial Difference LBP-TOP (RDLBPTOP), respectively. Hu *et al.* [80] combined LBP-TOP and learning based features to form a feature fusion for multi-task learning. Hu *et al.* [47] also took advantage of Gabor filter and proposed Centralized Gabor Binary Pattern from Three Orthogonal Panels (CGBP-TOP). Yu *et al.* [46] proposed a new Local Cubes Binary Patterns (LCBP) especially for ME spotting. Instead of using three different plane combinations to represent the spatio-temporal feature [82], [83], they utilized a cube mask to encode the information of eight directional angles for both the space and time domains.

Gradient-Based Feature. One of the key challenges for the feature descriptor is to describe the subtle changes in ME sequences. Aside from LBP, local patterns based on gradient have been used for the property that high order gradient would represent the detailed structure information of an image. Dalal and Triggs proposed the histogram of gradients (HOG) in 2005 [84], which is one of the most commonly used feature descriptors when it comes to object recognition for the ability to specialize the edges in an image. The HOG descriptor has the property of geometric invariance and optical invariance. With HOG descriptors, the expression contour feature can be well captured.

Histogram of Image Gradient Orientation (HIGO) is a variant of HOG. It ignores the magnitude weighting in the original HOG and thus can suppress the illumination effect. Zhang *et al.* [70] utilized both HOG-TOP and HIGO-TOP as the feature descriptors in their work of MER. Tran *et al.* [56] built a spotting network based on LSTM using HOG-TOP and HIGO-TOP features as the input of the network. To better capture the structural changes in ME videos using

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 3
Performance comparison among micro-expression spotting methods

Methods	Year	Features	Protocol	Datasets	Experimental Results		
					ACC	F1-score	Recall
[53]	2020	Specific Pattern	-	SAMM	0.090	0.133	0.258
				CAS(ME)^2	0.029	0.055	0.456
[54]	2019	HOOF&RNN	LOSO	SAMM	0.045	0.082	0.465
[55]	2019	LTP-ML	LOSO	SAMM	0.017	0.032	0.296
				CAS(ME)^2	0.009	0.018	0.281
[56]	2019	HIGO-TOP & HOG-TOP	LOSO	SMIC-VIS-E	-	0.620	-
				CASME II	-	0.860	-
[57]	2019	HOG	LOSO	CASME II	0.823	-	-
[58]	2019	time-constrasted feature	-	CASME II	0.815	-	-
[59]	2018	collaborative feature difference (CFD)	-	SMIC-E	-	-	0.942(AUC)
				CASME II	-	-	0.971(AUC)
[60]	2018	reiesz pyramid	LOSO	SMIC-HS	-	-	0.898(AUC)
				CASME II	-	-	0.951(AUC)

TABLE 4
Performance comparison among *macro-to-micro adaptation* methods

Methods	Year	Macro Datasets	Features	Protocol	Micro Datasets	Experimental Results		
						ACC	F1-score	UAR
[71]	2020	CK+	ResNet18	LOSO (cross-dataset)	CASME II	0.756	0.701	0.872
		MMI			SAMM	0.741	0.736	0.819
		Oulu-CASIA			SMIC	0.768	0.744	0.861
[72]	2019	Oulu-CASIA	3D CNN	5-fold cross validation	CASME II	0.976	-	-
					SAMM	0.974	-	-
[73]	2019	BU-3DFE	ResNet	LOSO (cross-dataset)	CASME II	-	0.761	0.755
					SAMM	-	0.448	0.487
					SMIC	-	0.551	0.546
[74]	2018	CK+	LBP-TOP	-	CASME II	0.836	0.856	0.863
[75]	2018	CK+	CNN	LOSO	CASME II	0.659	0.539	0.584
		Oulu-CASIA			SAMM	0.485	0.402	0.559
		Jaffe, MUGFE			SMIC	0.494	0.496	-
[76]	2018	CK+, Oulu-CASIA	ResNet10	LOSO (cross-dataset)	CASME II	0.757	0.650	-
		Jaffe, MUGFE			SAMM	0.706	0.540	-
[77]	2018	CK+	DCP-TOP & HWP-TOP	-	CASME II	0.607	-	-

gradient information, Niu *et al.* [49] proposed a local pattern of Local Two-Order Gradient Pattern (LTOGP).

Optical Flow Based Feature. The feature descriptors based on optical flow infer the relative motion information between different frames in order to capture subtle muscle movements for MER. The idea of optical flow was first introduced by Horn *et al.* [85] to describe the movement of brightness patterns in an image. By utilizing the pixel-wise difference between consecutive frames in a video clip, the motion information of the object in the video can thus be obtained. The basic concept is to find the distance of an identical object in different frames.

Owing to the fact that optical flow could capture temporal patterns between consecutive frames, one of the most employed architecture is to combine optical flow feature with CNN to further recognize spatial patterns [86], [87], [88], [89], [90], [91].

Verburg *et al.* [54] utilized Histogram of Oriented Optical Flow (HOOF) to encode the subtle changes in the time domain for selected face regions. Li *et al.* [92], [93] revisited the HOOF feature descriptor and proposed an enhanced version to reduce the redundant dimensions in HOOF. Liong *et al.* [50] proposed another feature descriptor based on optical flow, Bi-Weighted Oriented Optical Flow (Bi-WOOF). Bi-WOOF represents a sequence of subtle expressions using only two frames. In contrast to HOOF, both the magnitude and optical strain values are used as weighting schemes to highlight the importance of each optical flow so the noisy optical flows with small intensities are reduced.

However, using histogram of optical flow as feature descriptors may have some flaws. When the histogram is used as a feature vector for a classifier, even a slightly shifted version of a histogram will create a huge difference as most classification algorithms use euclidean distance to measure the difference of two images. Happy *et al.* [94] proposed Fuzzy Histogram of Optical Flow Orientations (FHOFO) to collect the motion directions into angular bins based on the

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

fuzzy membership function. It ignores the subtle motion magnitudes during the MEs and only takes the motion direction into consideration.

Other Handcrafted Feature. Apart from the abovementioned features, there are other descriptors to capture the distinctive properties embedded in ME videos. Please refer to our supplementary materials for reference.

4.2 Learning based Feature

In recent years, deep learning has shown a substantial improvement in computer vision tasks [98], [99], [100], [101], [102], [103], [104], [105]. Convolutional Neural Networks (CNNs, or ConvNet) have been considered as the most widely used ways for visual feature extraction with discriminative ability [98].

Since MEs are sequences of images, 2D CNN and 3D CNN are useful for their representative feature learning. 2D CNN is generally used on image data, which was first introduced in LeNet5 [106]. It is called two dimensional CNN because the kernel slides along two dimensions on the data. Input and output data in 3D CNN are four-dimensional since the kernel moves in three directions. 3D CNN is mostly used on 3D image data, e.g. MRI, CT scans, and videos [107], [108], [109].

Shallow learning. Deep learning-based methods require large-scale data to train the model. However, current public ME datasets are mostly limited and imbalanced, and tend to cause over-fitting issues when directly applying ConvNet. Most works usually adopt shallow and lightweight layers for MER. Dual Temporal Scale Convolutional Neural Network (DTSCNN) [110] was the first end-to-end middle-size neural network for MER. The DTSCNN designs two temporal channels, where each channel only has 4 convolutional layers and 4 pooling layers to partially avoid over-fitting. Peng *et al.* [76] adopted a simplified version of ResNet, ResNet10 for representative learning. Takalkar *et al.* [111] employed five convolutional layers, three max-pooling layers, and three fully-connected layers. Zhao *et al.* [90] used four convolutional layers and three pooling layers for capturing discriminative and high-level ME features. The 1×1 convolutional layer is added right after the input layer to increase the non-linear expression of input data without increasing the computational load of the model. Micro-attention [75] was built with 10 residual blocks, whose micro-attention unit is an extension of ResNet. The shortcut connection designed for identity mapping can reduce the degradation problem. By learning the spatial attention of feature maps, the network can focus on the facial subtle movements.

Two-step learning. There are many works that first extract spatial features among all frames, then use recurrent convolutional layer or LSTM module to explore their temporal correlation. Due to the small amount of training samples, many learning based works [86], [113], [54] include handcrafted features (e.g., optical flow, HOOF) to give a higher signal-to-noise (SNR) ratio in comparison to using raw pixel data.

The optical flow map and its related extension algorithms, e.g., HOOF, are commonly used for extracting motion deformations of facial regions and have shown good performance [54], [86], [113], [54]. RCN-F [86] first extracted an optical flow map from the located onset and apex frames followed by a recurrent convolutional network (RCN). The RCN is composed of three parameter-free modules (i.e., wide expansion, shortcut connection, and attention unit) to enhance the representational ability. Enriched Long-term Recurrent Convolutional Network (ELRCN) [113] also used the optical flow as input data for two learning modules: dimension enrichment and temporal dimension enrichment.

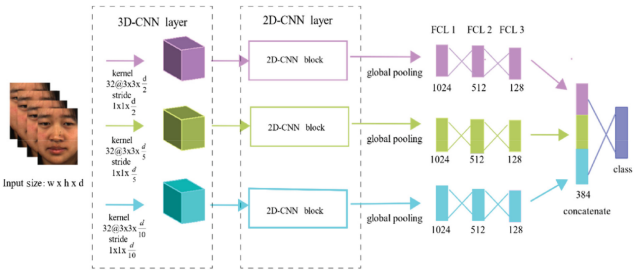

Fig. 6. The combination structure of the 2D CNN and 3D CNN in TSNN-LF for MER [112].

3D CNN. In 2D CNN, convolution and pooling operations are applied on 2D feature maps which is lacking of motion information. To better preserve temporal information of the input signals effectively, 3D convolution has been proved to have strong ability on capturing more representative features in the spatio-temporal aspect [107]. Considering that MEs occur only in a short period among dynamic facial movements and can be explicitly analyzed by detecting their constituent temporal segments, several MER works [114], [115], [72], [116], [112] used 3D CNN to extract the facial temporal movements.

Li *et al.* [115] applied the 12-layer 3D flow-based CNNs model for MER, which extracts motion flow information arising from subtle facial movements. To better represent the dynamic and appearance features of MEs, optical flow and gray scale frames are combined as input data. The 3D convolution adopts 3×3×3 kernels to represent the spatial changes at each local region. Zhi *et al.* [72] employed 3D CNN for self-learning feature extraction. All convolutional kernels are set to the size of 3×5×5, where 3 is the temporal depth and 5×5 is the size of the spatial receptive field by taking consideration of efficiency and computational complexity. Reddy *et al.* [116] proposed two 3D CNN models for spontaneous MER, i.e., MicroExpSTCNN and MicroExpFuseNet, by exploiting the spatio-temporal information in the CNN framework.

Nevertheless, applying deep 3D CNN from scratch significantly increases the computational cost and memory demand. Besides, most popular 3D networks, e.g., C3D and P3D ResNet, are trained on Sports 1-M, that is an action database and is very different from MEs. Hence, it still needs to investigate the trained deep neural networks for ME feature extraction, which would boost the performance of the MER task [69]. TSNN [112] designed a three-stream combining 2D and 3D CNN to extract expression sequence feature. The 3D kernels were made up to have same spatial size and different lengths in time domain to avoid the use of pooling layer and thus more informative features could be reserved to improve the MER performance. The architecture

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 5
Performance comparison among *recognition based on key apex frames* methods

Methods	Year	Key Frames	Features	Protocol	Datasets	Experimental Results		
						ACC	F1-score	UAR
[95]	2020	Key Frame (SSIM)	dual-cross patterns (DCP)	-	CASME II	0.687	-	-
[86]	2020	Onset & Apex	RCN	LOSO	CASME II	-	0.856	0.812
					SAMM	-	0.765	0.677
					SMIC	-	0.658	0.660
[96]	2019	OFF & Apex	optical flow & CNN	LOSO	CASME II*	0.883	0.870	-
					SAMM*	0.682	0.542	-
					SMIC*	0.677	0.671	-
[90]	2019	Onset & Apex	optical flow	-	CASME II	0.870	-	-
					SAMM	0.698	-	-
					SMIC	0.702	-	-
[97]	2018	Apex	Bi-WOOF	-	CASME II	0.589	0.610	-
					SAMM	0.622	0.620	-
[73]	2019	Apex	ResNet	LOSO (cross-dataset)	CASME II	-	0.761	0.755
					SAMM	-	0.448	0.487
					SMIC	-	0.551	0.546
[76]	2018	Apex	ResNet10	LOSO (cross-dataset)	CASME II	0.757	0.650	-
					SAMM	0.706	0.540	-

* on upper-right of the name of datasets means the labels of positive, negative and surprise were used in the experiments.

of TSNN is shown in Fig. 6.

5 Micro-expression Spotting

Micro-expression spotting refers to locating segments of micro movements for a given video. There are less publications compared to MER in the literature due to the difficulty of discovering subtle difference in a short duration (see Fig. 1). However, ME spotting is a vital step for automatic ME analysis since proper spotting can decrease the redundant information for further recognition. Table 3 compares the *state-of-the-art* micro-expression spotting methods.

Tran *et al.* [54] first introduced deep sequence model for ME spotting. LSTM shows its effectiveness on using both local and global correlations of the extracted features to predict the score of the ME apex frame. Li *et al.* [25] first proposed an ME spotting method which was effective on spontaneous ME datasets. A spotting ME convolutional network [117] was designed for extracting features from video clips, which is the first time that deep learning is used in ME spotting.

The co-occurrence of macro- and micro-expressions are common in real life. An automatic spotting system for micro- and macro-expressions was designed by [53]. A new spotting benchmark has been proposed recently [118].

Onset and offset detection. While it is relatively easier to identify the peaks and valleys of facial movements, the onset and offset frames are much more difficult to determine. Locating the onset and offset frames is crucial for real-life situations where facial movements are continuously changing. CASME II and SAMM provide the onset and offset frame ground-truth, which can be helpful for model training.

Apex frame spotting. Liong *et al.* [119] introduced an automatic apex frame spotting method and this strategy was also adopted in [73], [97], [91]. The LBP feature descriptor was first employed to encode the features of each frame, then a *divide-and-conquer* methodology was exploited to detect the frames with peak facial changes as apex frame.

Most of the aforementioned ME spotting methods were conducted on public ME lab-controlled datasets. However, ME spotting in real-world scenes with different environmental factors is still an open issue. More ME datasets containing long video and real-world scenes are essential for further research.

6 MER Methodology

In this section, we aim to cover the recognition approaches, especially with deep learning techniques, which are not fully discussed in other survey papers. Considering the difficulties of MER mainly due to the lack of labeled data, subtle duration and low intensity, many works seek help from macro-expression data or take advantage of the most expressive frames. Moreover, AUs analysis is crucial to eliminate the ambiguity of emotion states. Hence, we review them in three novel aspects: Macro-to-micro, Recognition based on Key Apex Frames and Recognition based on Facial Action Units.

6.1 Macro-to-Micro Adaptation

As mentioned in the previous section, how to solve automatic labeling and recognition problems for MEs is a challenging task under the condition of a small number of labeled training ME samples available. Meanwhile, there are large amounts of macro-expression databases [120], [121], each of which consists of vast labeled training samples compared with micro-expression databases. Thus, how to take advantage of the macro-expression databases for MER has become an important direction for research. Table 4 summarizes the *macro-to-micro adaptation* methods.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

Transfer-learning based method has proved to be efficient in applying deep CNN on small databases [122]. Thus, by using the idea of transfer learning, it is reasonable to take advantage of the quantitative superiority of macro-expression to recognize the micro-expression [76], [74], [75], [72], [73]. In [74], the macro-expression features and micro-expression features are considered gallery and probe features respectively and constructed as training matrix, where the gallery features can be transformed to the probe features. The two different features were then projected into a joint subspace, where they are associated with each other. The nearest neighbor (NN) classifier was used to classify the probe micro-expression samples at the last step. Zhi *et al.* [75] pretrained the 3D CNN on macro-expression database Oulu-CASIA [123], and then the pre-trained model was transferred to the target micro-expression domain. The experimental results show 3.4% and 1.6% in MER performance higher than the model without transfer learning, respectively.

Xia *et al.* [71] imposed a loss inequality regularization to make the output of MicroNet converge to that of MacroNet. In [77], macro-expression images and micro-expression sequences were encoded by proposed hot wheel patterns (HWP), Dual-cross patterns (DCP-TOP) and HWP-TOP, respectively. The coupled metric learning algorithm was employed to model the shared features between micro-expression samples and macro-information. In [75], the original residual network (without micro-attention units) was first initialized with the ImageNet database [98]. Then, to narrow the gap between object recognition (ImageNet) and facial expression recognition, the network was further pre-trained on several popular macro-expression databases, including CK+ [124], Oulu-CASIA NIR & VIS [123], Jaffe [125], and MUGFE [126]. Finally, the residual network together with micro-attention units is fine-tuned with micro-expression databases, including CASME II, SAMM and SMIC. Similarly, Jia *et al.* [76] proposed a macro-to-micro transformation network. ResNet10 pre-trained on ImageNet dataset was fine-tuned on macro-expression datasets first and then on the micro-expression datasets (CASME II, SAMM).

Knowledge distillation strategy is also used for MER. Sun *et al.* [51] proposed a multi-task teacher network containing AUs recognition and facial view classification on FERA2017 dataset [127], then the learned knowledge was distillated to a shallow student network for MER. In SA-AT [73], the larger scale of macro-expression data were served as auxiliary database to train a teacher model, then the teacher model was transferred to train the student model on micro-expression databases with limited samples. To narrow the gap between macro- and micro-expressions, a style aggregated strategy was used to transform micro-expression samples from different macro-expression datasets to generate an aggregated style via CycleGAN [128].

6.2 Recognition based on Key Apex Frames

Apex frame, which is the instant indicating the most expressive emotional state in a video sequence, is effective to classify the emotion in that particular frame. The apex frame portrays the highest intensity of facial motion among all frames. The apex occurs when the change in facial muscle reaches the peak or the highest intensity of the facial motion. Many works [96], [86] extracted features of the apex frame as feature descriptors.

Existing ME datasets, e.g., CASME II and SAMM, provide the index annotation of apex frames, while SMIC-HS database does not release the apex information. Thus, automatic apex frame spotting [119], [73], [97], [91] is necessary to be applied to each video to acquire the location of the apex frame. Table 5 summarizes the *state-of-the-art* key apex frames based MER approaches.

Using single apex frame. Considering the subtle motion change of ME frames in video, several works emphasized the use of a single apex frame to minimize the redundancy of the repetitive input frames. Using apex frame to represent the whole ME sequence was expected to reduce the computational complexity for feature learning [76], [91], [97], [73]. The experimental results of [51] showed that the effect of only using apex frame is better than onset-apex-offset sequence and the whole video.

The optical flows from the onset and apex frames are commonly used for extracting motion deformations of facial regions and can achieve good performance in subject independent evaluation [141], [87], [90], [142]. Zhou *et al.* [134] took the mid-position frames in ME samples based on the observation that most apex frames existing in the middle segment of a data sequence. OFF-ApexNet [96] adopted onset and apex frames to represent the ME details and extracted optical flow to encode the motion flow features based on two chosen frames. The two streams are combined in a fully-connected layer for MER. The similar strategy has been used in [86]. The obtained onset and apex frames are combined with the corresponding optical flow map as input of the proposed MER model. Zhao *et al.* [90] used the Total Variation-L1 (TV-L1) optical flow algorithm to calculate the motion information between the onset frame and the apex frame. Then, the corresponding optical flow feature map is obtained and fed into subsequent deep network learning.

Others. Key frames of each local region are computed by SSIM in [95], and then the dual-cross patterns (DCP) are applied to get the final feature vectors.

6.3 Recognition based on Facial Action Units

As physical subtle variations in facial expressions, AUs analysis are crucial because they constitute essential signals to be understood. Emotion states may be conceptually similar in terms of the difference in facial muscle movements (e.g., fear and disgust). There are many works focusing on facial AU recognition [143], [144], including AU occurrence detection [145], [146], [147], [148] and AU intensity estimation [149], [150]. Some surveys compare recent AUs recognition works, e.g., [151], [152].

Since the occurrence of AUs are strongly correlated, AU detection is usually considered as a multi-label learning problem. Several works considered the relationships among AUs and modeled AU interrelations to improve recognition accuracy [153], [154], [155]. However, most works rely on probabilistic graphical models with manually extracted features [156], [157]. Given that the graph has the ability of handling multi-relational data [155], Liu *et al.* [153] proposed

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 6
Performance comparison on CASME II dataset

Methods	Year	Features	Classifier	Protocol	Experimental Results		
					ACC	F1-score	UAR
[79]	2019	ELBPTOP	SVM	LOSO	0.739	0.690	-
[79]*	2019	ELBPTOP	SVM	LOSO	0.796	0.660	-
[47]	2019	CGBP-TOP	Gaussian kernel function	LOSO	0.658	-	-
[92]	2019	revised HOOF + CNN	MLP	LOSO	0.580	-	-
[129]	2019	LMP	SVM	LOSO	0.702	-	-
[130]	2019	spatiotemporal texture map	SVM	k-fold cross validation	0.800	0.890	-
[131]*	2019	landmark point feature	MLP	LOSO	0.793	-	-
[52]	2018	discriminative STLBP	SVM + Gentle adaboost	LOSO	0.648	-	-
[132]	2018	FMBH	SVM	LOSO	0.643	-	-
[133]	2018	hierarchical ST descriptor	GSL model	LOSO	0.639	0.613	-
[86] *	2020	optical flow + RCN	MLP	LOSO	-	0.856	0.813
[87]*	2019	optical flow+STSTNet	MLP	LOSO	-	0.680	0.701
[88]*	2019	optical flow + CNN	MLP	LOSO	0.883	0.897	-
[89]*	2019	optival flow + CNN	MLP	LOSO	-	0.829	0.820
[134]*	2019	optical flow + CNN	MLP	LOSO	-	0.862	0.856
[135]	2019	LBP-TOP + optical flow	SVM	LOSO	0.696	-	-
[90]	2019	optical flow + CNN	MLP	LOSO	0.870	-	-
[136]	2019	dynamic imaging + CNN	MLP	k-fold cross validation	0.766	-	-
[95]	2019	local ROI+DCP	chi-square distance	-	0.687	-	-
[80]	2018	LGBP-TOP + CNN	MLP	LOSO	0.662	-	-
[137]	2018	frame difference + ST Gabor filter	SVM	LOSO	0.553	-	-
[138]	2021	CNN	Softmax	LOSO	0.942	-	-
[114]	2021	3D CNN	Softmax	LOSO	-	0.882	0.876
[139]	2021	CNN	Softmax	LOSO	0.848	0.848	-
[35]	2020	3D CNN	Softmax	LOSO & LOVO	0.561	0.394	-
[140]	2020	Graph-TCN	Softmax	LOSO	0.740	0.725	-
[71]	2020	ResNet18	Triplet loss	LOSO	0.756	0.701	0.872
[115]	2019	3D flow-based CNN	SVM	LOSO	0.591	-	-
[72]	2019	3D CNN	SVM	5-fold cross validation	0.976	-	-
[76]	2018	ResNet10	-	LOSO	0.757	0.650	-
[75]	2018	CNN	Softmax	LOSO	0.659	0.539	0.584

The * on upper-right of the methods mean only three labels of positive, negative and surprise were used in the experiments .

the first work that employed GCN to model AU relationship. The cropped AU regions by EAC-Net [148] are fed into GCN as nodes, after that the propagation of the graph is determined by the relationship of AUs.

AU intensities are annotated by appending five-point ordinal scale (A–E for minimal-maximal intensity) [152]. The intensity levels are not uniform intervals. To infer the co-occurences of AUs, it can be formulated into a heatmap regression-based framework. Fan *et al.* [149] modeled AU co-occurring patterns among feature channels, where semantic descriptions and spatial distributions of AUs are both encoded.

Further insights into locating AUs could possibly provide even better discrimination between types of MEs. Some public ME datasets, e.g., CASME II and SAMM, contain emotion classes as well as AUs annotation, which makes the relationship construction between ME classes and AUs become achievable. Until now, there is few literature exploring automatic MER based on AUs [14], [158], [154], [159]. Wang *et al.* [14] defined 16 ROIs based on AUs and features extracted for each ROI can further boost the MER performance. Liu *et al.* [158] partitioned the facial region into 36 ROIs from 66 feature points and proposed a ROI-based optical feature, MDMO for MER. However, such predefined rules for modeling relationships among AUs may lead to limited generalization. Li *et al.* [154] used a structured knowledge-graph for AUs and integrated a Gated GNN (GGNN) to generate enhanced AU representation. This is the first work that integrates AU detection with MER. Lo *et al.* [159] proposed MER-GCN, which was the first end-to-end AU oriented MER architecture based on GCN, where GCN layers are able to discover the dependency laying between AU nodes for MER. Graph-TCN [140] utilized the graph structure for node and edge feature extraction, where the facial graph construction is shown in Fig. 7. Sun *et al.* [51] proposed a knowledge transfer strategy that distills and transfers AUs information for MER.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 7
Performance comparison on SMIC dataset

Methods	Year	Features	Classifier	Protocol	Experimental Results		
					ACC	F1-score	UAR
[79]	2019	ELBPTOP	SVM	LOSO	0.691	0.650	0.660
[47]	2019	CGBP-TOP	Gaussian kernel function	LOSO	0.594	-	-
[129]	2019	LMP	SVM	LOSO	0.674	-	-
[52]	2018	discriminative STLBP	SVM + Gentle adaboost	LOSO	0.634	-	-
[132]	2018	FMBH	SVM	LOSO	0.683	-	-
[50]	2018	Bi-WOOF	SVM	LOSO	0.620	-	-
[94]	2018	FHOFO	SVM	LOSO	0.512	0.518	-
[133]	2018	hierarchical ST descriptor	GSL model	LOSO	0.604	0.613	-
[86]	2020	optical flow + RCN	MLP	LOSO	-	0.658	0.660
[135]	2019	LBP-TOP + optical flow	SVM	LOSO	0.717	-	-
[87]	2019	optical flow + STSTNet	MLP	LOSO	-	0.659	0.681
[89]	2019	optical flow + CNN	MLP	LOSO	-	0.746	0.753
[134]	2019	optical flow + CNN	MLP	LOSO	-	0.664	0.673
[88]	2019	optical flow + CNN	MLP	LOSO	0.677	0.671	-
[90]	2019	optical flow + CNN	MLP	LOSO	0.698	-	-
[136]	2019	dynamic imaging + CNN	MLP	k-fold	0.911	-	-
[80]	2018	LGBP-TOP + CNN	MLP	LOSO	0.651	-	-
[137]	2018	frame difference + ST Gabor filter	SVM	LOSO	0.545	-	-
[114]	2021	3D CNN	Softmax	LOSO	-	0.736	0.760
[139]	2021	CNN	Softmax	LOSO	0.787	0.797	-
[71]	2020	ResNet18	Triplet loss	LOSO	0.768	0.744	0.861
[115]	2019	3D flow-based CNN	SVM	LOSO	0.555	-	-
[116]	2019	3DCNN	Softmax	-	0.650	-	-
[75]	2018	CNN	Softmax	LOSO	0.494	0.496	-

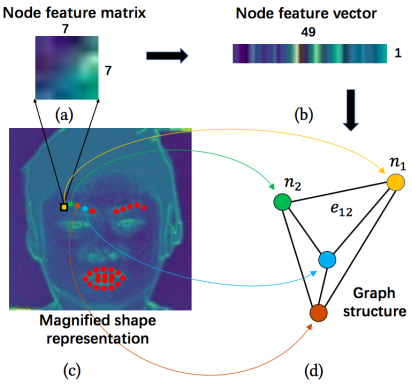

Fig. 7. Sample facial graph built in Graph-TCN [140].

7 Loss Function

7.1 Ranking Loss

Ranking Loss, such as *Contrastive Loss*, *Margin Loss*, *Hinge Loss* or *Triplet Loss*, is widely used in MER [70], [160], [136], [55], [161], [162], [71], [163]. The objective of ranking loss is to predict relative distances between inputs, which is also called metric learning [164]. Lim and Goh [163] proposed Fuzzy Qualitative Rank Classifier (FQRC) to model the ambiguity in MER task by using a multi-label rank classifier. Xia *et al.* [71] imposed an loss inequality regularization in triplet loss to make the output of MicroNet converge to that of MacroNet. The hinge loss is used for maximum-margin classification, most notably for Support Vector Machines (SVMs). SVM classify training data points by finding a discriminative hyperplane function according to different kernels such as Linear kernel, Polynomial kernel, and Radial Basis Function (RBF) kernel. Each different kernel has its own advantage for different separability and order of the dataset. Li *et al.* [55] split each video into 12 feature ensembles, which represent the ME local movements, and SVM was employed to classify these local ROIs. LEARNet [136] employed RankSVM [165] to compute the frame scores in a video. To reduce computation time, Sequential Minimal Optimization (SMO), one of the computationally fastest methods of evaluating linear SVMs, was used in [161].

Overall, SVM is one of the most well-applied classifiers for MER. However, they show poor performances when the feature dimension is far greater than the training sample number.

7.2 Cross-Entropy Loss

Unlike traditional methods, where the feature extraction step and the feature classification step are independent, deep networks can perform MER in an end-to-end way. Face recognition system, e.g., Deepface [166], first adopted cross-entropy based softmax loss for facial feature learning. Specifically, softmax loss is the most commonly used function that minimizes the cross-entropy between the estimated class probabilities and the ground-truth distribution. Most works directly applied softmax loss in the MER network [139], [114], [54], [86], [96]. For example, the softmax

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

TABLE 8
Performance comparison on SAMM dataset

Methods	Year	Features	Classifier	Protocol	Experimental Results		
					ACC	F1-score	UAR
[79]	2019	ELBPTOP	SVM	LOSO	0.634	0.480	-
[79]*	2019	ELBPTOP	SVM	LOSO	-	0.780	0.720
[131]*	2019	landmark point feature	MLP	LOSO	0.746	-	-
[86]*	2020	optical flow + RCN	MLP	LOSO	-	0.765	0.677
[87]*	2019	optical flow + STSTNet	MLP	LOSO	-	0.838	0.869
[89]*	2019	optical flow + CNN	MLP	LOSO	-	0.775	0.715
[134]*	2019	optical + CNN	MLP	LOSO	-	0.587	0.566
[88]*	2019	optical + CNN	MLP	LOSO	0.682	0.542	-
[90]	2019	optical + CNN	MLP	LOSO	0.702	-	-
[138]	2021	CNN	Softmax	LOSO	0.867	0.870	-
[114]	2021	3D CNN	Softmax	LOSO	-	0.748	0.728
[139]	2021	CNN	Softmax	LOSO	0.811	0.744	-
[71]	2020	ResNet18	Triplet loss	LOSO	0.741	0.736	0.819
[35]	2020	3D CNN	Softmax	LOSO & LOVO	0.523	0.357	-
[72]	2019	3D CNN	SVM	5-fold cross validation	0.974	-	-
[75]	2018	CNN	Softmax	LOSO	0.485	0.402	0.559

The * on upper-right of the methods mean only three labels of positive, negative and surprise were used in the experiments .

function was employed followed by an LSTM network which consists of two LSTM layers, each with 12 dimensions [54].

7.3 Others

Since the MER task suffers from not only high inter-class similarity but also high intra-class variation, instead of simply employed SVM or softmax loss, several works have proposed novel loss layers for MER [167], [168], [169], [170].

Inspired by the center loss, which penalizes the distance between deep features and their corresponding class centers, Lalitha *et al.* [167] combined cross-entropy loss and center loss to improve the discriminative energy of the deeply learned features. The focal loss [168] was utilized in MACNN [169] to overcome the sample imbalance challenge. Compared with cross-entropy loss, focal loss is more suitable to solve the problem of small sample classification or imbalance between samples [114]. Besides, Xie *et al.* [170] proposed a *feature loss* metric to use the complementary information of handcrafted and deep features during training.

8 Experiments

8.1 Evaluation

8.1.1 Evaluation Protocols

The commonly used evaluation protocols for MER are *k*-fold cross-validation, leave-one-subject-out (LOSO) and leave-one-video-out (LOVO). *k*-fold cross-validation repeats random sub-sampling validation. For LOSO validation, the model leaves out all samples of one single subject for the performance evaluation, and all other data are used as training data. The overall performance is then averaged from all folds. Similar to LOSO, LOVO validation protocol requires the model to pick up the frames from one video for validation purpose while all other data are sampled for training. From our review, the LOSO is the most widely used.

There are limitations for the above protocols. With severe class-imbalanced ME data, the effectiveness of *k*-fold is easily influenced. The same problem goes with LOVO, which can introduce additional biases on certain subjects that have more representations during the evaluation process. Moreover, the evaluation results may be over-estimated due to the large training data. For LOSO, the intrinsic ME dynamics of each subject may be limited since the intensity and manner of MEs may differ from person to person.

8.1.2 Evaluation Metrics

The popularly used MER evaluation metrics include Accuracy (ACC) and F1-score. ACC shows the average hit rate across all classes and is susceptible to bias data, and thus it can only reflect the partial effectiveness of an MER classifier. F1-score can remedy the bias issue by computing on the total true positives, false positives and false negatives to reveal the ME classes.

For cross-dataset evaluation, unweighted average recall (UAR) and weighted average recall (WAR) are commonly used [2], [171]. WAR refers to the number of correctly classified samples divided by the total number of samples. UAR is defined as mean accuracy of each class divided by the number of classes without consideration of samples per class, which can reduce the bias caused by the class imbalance.

8.2 Cross-dataset MER

Cross-dataset MER is one of the recently emerging while challenging problems in ME analysis. The training sets and testing sets are from different ME datasets, resulting in the inconsistency of feature distributions [172].

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

The cross-database recognition in Micro-Expression Grand Challenge (MEGC 2018) [2] used the CASME II and SAMM datasets to serve as the source and target ME database. Zong *et al.* [171] built a cross-database micro-expression recognition (CDMER) benchmark. There are two types of CDMER tasks. The TYPE-I is conducted between any two subsets of SMIC, i.e., HS, VIS, and NIR. And the TYPE-II uses the selected CASME II and one subset of SMIC, which is proved more difficult than TYPE-I [171]. The experiments showed the variant of Local Phase Quantization (LPQ), i.e., LPQ-TOP, and LBP with six intersection points (LBP-SIP) performed the best in terms of F1-score and ACC in TYPE-I and TYPE-II. Also, deep features perform rather poorly than most handcrafted features in the CDMER task. The possible reason may be the C3D was pre-trained on Sports1M and UCF101 datasets, whose samples are quite different from ME data. After finetuning the C3D on SMIC (HS), the performance of C3D features can be improved significantly. However, with a simple finetuning strategy based on the source database, it seems not enough for learning the database-invariant deep features.

The heterogeneous problem existing between source and target databases raises the level of difficulty of the CDMER task. For example, SMIC (NIR) is collected by a near-infrared camera, whose image quality is considerably different from images recorded by a high-speed camera (as in CASME II and SMIC (HS)) and visual camera (as in SMIC (VIS)). Li *et al.* [173] proposed the target-adapted least-squares regression (TALSR) to learn a regression coefficient matrix between source samples and the provided labels. The idea is that such coefficient matrix could also reflect the sample-label relation in the target database domain. Zhang *et al.* [172] proposed a structure of the super wide regression network (SWiRN) for unsupervised cross-database MER. The *state-of-the-art* domain adaptation methods [174], [175] can be further exploited to reduce the differences between source and target domains to improve the performance of CDMER methods.

8.3 Overall Comparison

A comparison of MER methodologies with handcrafted features and learning-based features is provided in Table 6, Table 7 and Table 8, respectively. Several variants of optical flow and LBP-TOP have also been proposed in recent years. Also, optical flow-related features are the most frequently used handcrafted features. The experimental results show that the combination of optical flow features and CNN can achieve fine recognition accuracy. Among all methods using handcrafted feature, LEARNet [136] achieved averagely the best performance, proving the effectiveness of dynamic imaging and the feasibility of using key frames to represent a video sequence.

Although there are many approaches using handcrafed features, we can observe that the trend of MER is changing. In several approaches, low-level handcrafted feature extraction is considered as a pre-stage before high-level feature extraction and fine performances can be obtained. On the other hand, although the end-to-end learning based approaches are reported to have higher performances, they are still restricted by limited labeled data compared to other classification tasks in the literature, encouraging the development of data augmentation or transfer learning works.

While the obtained results from most works are encouraging, there are still some restrictions. Since there is no standard evaluation protocol and class setting, it is hard to reach the conclusions which method performs the best for MER. For example, there are some works only considering three or four emotion labels (i.e., Positive, Negative, Surprise, and Others) [50], [92]. The reduction of emotion classes makes the MER task simpler but also introduces the class bias towards negative categories since there is only one positive category (i.e., Happiness). Another problem is that different methods have different dataset splitting or composite strategies [86], [176]. For example, Lai *et al.* [169] mixed CASME and CASME II and split the training/testing set as 70/30 percentage. While SMIC, CASME II, and SAMM were combined into a composite dataset in [176].

Generally, the problem of real-world automatic micro-expression recognition and spotting still remain challenging, since the existing datasets are collected under the well-controlled environment and the data diversity is very limited. The recognition performances are insufficient and there remain still many topics unexplored.

9 Challenges and Potential Solutions

9.1 In-the-wild Scenarios

Current MER researches have limited scenarios since existing publicly annotated datasets are all collected from lab-controlled environments. In the real-world scenario, more complex and natural emotion including various environmental factors, e.g., illumination interference, 3D head rotation, interaction or interrogation scenarios where more persons involved should be taken into account. Lai *et al.* [169] took a step forward to avoid excessive computation and established a real-time MER framework with 60 fps running on Intel Xeon 2.10 GHz CPU, 32 GB memory and Ubuntu 14.04 operating system. However, it was experimented on CASME and CASME II, which both are lab-controlled datasets. The MER in-the-wild still remains an open challenge.

Meanwhile, the privacy issue should be considered in high priority during data collection. Researchers should abide by the laws and regulations when collecting facial expression data, sign data collection terms with subjects, and fully consider personal privacy issues. Currently, public experiments are conducted with few clear ethical guidelines. The importance of constructing an appropriate ethics is of some urgency since not all researchers will converge on consistent standards for experimental research in grey areas [177]. Hence, it is still an open question how ethical dilemmas arise in the conduct of social research (e.g., induce emotion in unsuspecting people) and how they can be resolved.

9.2 Uncertainty Modeling

To figure out why the *state-of-the-art* methods reached a bottleneck in the recognition rate, several works introduced fuzzy set theory [178] into MER due to its ability to model the uncertainties. For ambiguous movements of different

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

emotions, Fuzzy Qualitative can model the ambiguity by using the fuzzy membership function to represent each feature dimension with respect to each emotion class. Lim and Goh [163] considered ME as a non-mutual exclusive case. Instead of conducting crisp classification, Fuzzy Qualitative Rank Classifier (FQRC) was proposed to model the ambiguity in MER task by using a multi-label rank classifier. Chen *et al.* [179] proposed weighted fuzzy classification for analyzing emotion and achieved promising accuracy. Happy and Routray [94] constructed fuzzy histograms of orientation features based on optical flow.

9.3 Machine Bias

Previous studies have indicated that ingroup advantage in macro-expression recognition exists in various kinds of social groups, e.g., cultural groups, racial groups, and religious groups [180], [181], [182], [183], [184], [185], [186].

However, it remains unclear whether the social category of the target influences MER. Xie *et al.* [187] conducted the intergroup bias experiments among Chinese. The results showed that there is an ingroup disadvantage for the Chinese participants: The recognition accuracy of MEs of outgroup members (White targets) was actually higher than that of ingroup members (Asian targets). And the results also showed that such an intergroup bias is unaffected by the duration of MEs and the ingroup disadvantage remains the same even after the participants had received the training of Micro Expression Training Tool (METT). Regardless of the duration of MEs, there is an ingroup disadvantage for Chinese participants and such a bias still exists even after the training of MER training program.

One possible solution is to discover the hidden group-invariant features. Investigating the potential factors that may affect the MER help us to develop efficient MER training programs and efficient automatic recognition tools.

10 Conclusions and Future Research Directions

The past decade has witnessed the development of many new MER algorithms. This paper provides a comprehensive review of recent advances in MER technology. Future MER works can expand from certain research directions below.

10.1 Enriching Limited and Biased Datasets

Current ME databases are usually collected from young subjects, especially undergraduate students. The age range should be extended in the future because MEs decoding is varying among different ages [188]. Meanwhile, the male/female percentage and ethnic groups should also be well considered. Also, to reduce the labeling process, synthetic data generation algorithms can be further exploited in MER as introduced in Section 2.

Due to the subjectivity of human annotators and the ambiguous nature of the expression labels, there might exist annotation inconsistency [189], [190], how to reduce label noises is also needs to consider.

10.2 Facial Asymmetrical Phenomenon for MER

Chirality is a chemistry term used to describe two objects that appear identical but not symmetrical when folded over onto themselves. Since human faces are naturally asymmetrical [191], *facial chirality* demonstrates the asymmetry as the left and the right side of faces differ in emotional communication when people experience multiple emotions at the same time or when there is an attempt to hide an emotion. The general observation is that emotional expressions are more intense on the left side of faces since the right cerebral hemisphere is dominant for the expression [192]. On the other hand, AUs may be coded as symmetrical or asymmetrical. Some existing datasets have left and right AUs annotation that can reveal such phenomenon. Hypothetically, the facial chirality implies that the left side of faces might already include sufficient features that can distinguish one emotion from another. However, there are few MER works extend their researches based on this hypothesis.

10.3 Multimodality for MER

Auxiliary metadata, e.g., words, gestures, voices, can serve as important cues for MER in real world scenarios. Among them, the body gesture is proved to be capable of conveying emotional information. A *Micro Gesture* dataset [193] containing 3,692 manually labeled gesture clips was released in 2019. The dataset collects subtle body movements that are elicited when hidden expresions are triggered in unconstrained situations. Their experiments verified the latent relation between one's micro-gestures and the hidden emotional states. Therefore, MEs can be fused with micro-gesture and other physiological signals for a finer level emotion understanding. Moreover, human physiological features, e.g., ECG, EDA, are very informative features for affective revealing, which are less considered in MER works.

Acknowledgments

This work was supported in part by the CTBC Bank under Industry-Academia Cooperation Project and the Ministry of Science and Technology of Taiwan under Grants MOST-109-2223-E-009-002-MY3 and MOST-110-2218-E-A49-018.

References

[1] P. Ekman, *Telling lies: Clues to deceit in the marketplace, politics, and marriage (revised edition)*. WW Norton & Company, 2009.

[2] W. Merghani, A. Davison, and M. Yap, "Facial micro-expressions grand challenge 2018: evaluating spatio-temporal features for classification of objective classes," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2018, pp. 662–666.

[3] J. See, M. H. Yap, J. Li, X. Hong, and S.-J. Wang, "Megc 2019– the second facial micro-expressions grand challenge," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–5.

[4] J. Li, S. Wang, M. H. Yap, J. See, X. Hong, and X. Li, "Megc2020- the third facial micro-expression grand challenge," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2020, pp. 234–237.

[5] P. Ekman, "Lie catching and microexpressions," *The Philosophy of Deception*, vol. 1, no. 2, p. 5, 2009.

[6] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry*, vol. 32, no. 1, pp. 88–106, 1969.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

[7] X.-b. Shen, Q. Wu, and X.-l. Fu, "Effects of the duration of expressions on the recognition of microexpressions," *Journal of Zhejiang University Science B*, vol. 13, no. 3, pp. 221–230, 2012.

[8] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE transactions on Affective Computing*, vol. 9, no. 1, pp. 116–129, 2016.

[9] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, and X. Fu, "CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–7.

[10] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, and X. Fu, "CASME II: An improved spontaneous micro-expression database and the baseline evaluation," *PloS One*, vol. 9, no. 1, p. e86041, 2014.

[11] X. Shen, Q. Wu, K. Zhao, and X. Fu, "Electrophysiological evidence reveals differences between the recognition of microexpressions and macroexpressions," *Frontiers in Psychology*, vol. 7, p. 1346, 2016.

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

[13] E. L. Rosenberg and P. Ekman, *What the face reveals: Basic and applied studies of spontaneous expression using the facial action coding system (FACS)*. Oxford University Press, 2020.

[14] S.-J. Wang, W.-J. Yan, X. Li, G. Zhao, C.-G. Zhou, X. Fu, M. Yang, and J. Tao, "Micro-expression recognition using color spaces," *IEEE Transactions on Image Processing*, vol. 24, no. 12, pp. 6034–6047, 2015.

[15] M. Pantic, "Machine analysis of facial behaviour: Naturalistic and dynamic behaviour," *Philosophical Transactions of the Royal Society B: Biological Sciences*, vol. 364, no. 1535, pp. 3505–3513, 2009.

[16] W. V. Friesen and P. Ekman, *EMFACS-7: Emotional Facial Action Coding System, Version 7*. Unpublished manual, 1984.

[17] P. Ekman, W. V. Friesen, and J. C. Hager, "Facial action coding system: The manual on cd rom," *A Human Face, Salt Lake City*, pp. 77–254, 2002.

[18] R. Cairns, G. Elder Jr, and E. Costello, "Cambridge studies in social and emotional development," *Developmental science. Cambridge University Press*, 1996.

[19] J. I. Durán, R. Reisenzein, and J.-M. Fernández-Dols, "Coherence between emotions and facial expressions," *The science of facial expression*, pp. 107–129, 2017.

[20] X. Li, T. Pfister, X. Huang, G. Zhao, and M. Pietikäinen, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2013, pp. 1–6.

[21] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, and X. Fu, "CAS (ME) ^{2}: A database for spontaneous macro-expression and micro-expression spotting and recognition," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 424–436, 2017.

[22] C. H. Yap, C. Kendrick, and M. H. Yap, "SAMM long videos: A spontaneous facial micro-and macro-expressions dataset," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020, pp. 194–199.

[23] X. Ben, Y. Ren, J. Zhang, S.-J. Wang, K. Kpalma, W. Meng, and Y.-J. Liu, "Video-based facial micro-expression analysis: A survey of datasets, features and algorithms," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021.

[24] W.-J. Yan, S.-J. Wang, Y.-J. Liu, Q. Wu, and X. Fu, "For micro-expression recognition: Database and suggestions," *Neurocomputing*, vol. 136, pp. 82–87, 2014.

[25] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, and M. Pietikäinen, "Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods," *IEEE Transactions on Affective Computing*, vol. 9, no. 4, pp. 563–577, 2017.

[26] K. M. Goh, C. H. Ng, L. L. Lim, and U. U. Sheikh, "Micro-expression recognition: an updated review of current trends, challenges and solutions," *The Visual Computer*, vol. 36, no. 3, pp. 445–468, 2018.

[27] Y.-H. Oh, J. See, A. C. Le Ngo, R. C.-W. Phan, and V. M. Baskaran, "A survey of automatic facial micro-expression analysis: databases, methods, and challenges," *Frontiers in Psychology*, vol. 9, p. 1128, 2018.

[28] W. Merghani, A. K. Davison, and M. H. Yap, "A review on facial micro-expressions analysis: datasets, features and metrics," *arXiv preprint arXiv:1805.02397*, 2018.

[29] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: facial micro-expression recognition," *Multimedia Tools and Applications*, vol. 77, no. 15, pp. 19 301–19 325, 2018.

[30] L. Zhou, X. Shao, and Q. Mao, "A survey of micro-expression recognition," *Image and Vision Computing*, p. 104043, 2020.

[31] L. Zhang and O. Arandjelović, "Review of automatic microexpression recognition in the past decade," *Machine Learning and Knowledge Extraction*, vol. 3, no. 2, pp. 414–434, 2021.

[32] P. Ekman, "Facial expressions of emotion: an old controversy and new findings," *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, vol. 335, no. 1273, pp. 63–69, 1992, publisher=The Royal Society London.

[33] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3d-gradient descriptor," 2009.

[34] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2011, pp. 51–56.

[35] H.-X. Xie, L. Lo, H.-H. Shuai, and W.-H. Cheng, "Au-assisted graph attention convolutional network for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2871–2880.

[36] M. Frank, M. Herbasz, K. Sinuk, A. Keller, and C. Nolan, "I see how you feel: Training laypeople and professionals to recognize fleeting emotions," in *The Annual Meeting of the International Communication Association*. ICA, 2009, pp. 1–35.

[37] G. J. Qi and J. Luo, "Small data challenges in big data era: A survey of recent progress on unsupervised and semi-supervised methods," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2020.

[38] R. Queiroz, M. Cohen, J. L. Moreira, A. Braun, J. C. J. Júnior, and S. R. Musse, "Generating facial ground truth with synthetic faces," in *IEEE Conference on Graphics, Patterns and Images*. IEEE, 2010, pp. 25–31.

[39] 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 *IEEE International Conference on Computer Vision Workshops*. IEEE, 2017, pp. 1609–1618.

[40] H. Ding, K. Sricharan, and R. Chellappa, "Exprgan: Facial expression editing with controllable expression intensity," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, 2018.

[41] R. Wu, G. Zhang, S. Lu, and T. Chen, "Cascade EF-GAN: Progressive facial expression editing with local focuses," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 5021–5030.

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

[43] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly, and Y. Tong, "Identity-free facial expression recognition using conditional generative adversarial network," *arXiv preprint arXiv:1903.08051*, 2019.

[44] A. Kortylewski, B. Egger, A. Schneider, T. Gerig, A. Morel-Forster, and T. Vetter, "Analyzing and reducing the damage of dataset bias to face recognition with synthetic data," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, 2019, pp. 0–0.

[45] T. Schlett, C. Rathgeb, O. Henniger, J. Galbally, J. Fierrez, and C. Busch, "Face image quality assessment: A literature survey," *arXiv preprint arXiv:2009.01103*, 2020.

[46] M. Yu, Z. Guo, Y. Yu, Y. Wang, and S. Cen, "Spatiotemporal feature descriptor for micro-expression recognition using local cube binary pattern," *IEEE Access*, vol. 7, pp. 159 214–159 225, 2019.

[47] C. Hu, J. Chen, X. Zuo, H. Zou, X. Deng, and Y. Shu, "Gender-specific multi-task micro-expression recognition using pyramid cgbp-top feature," *Computer Modeling in Engineering and Sciences*, vol. 118, no. 3, pp. 547–559, 2019.

[48] P. Carcagnì, M. Del Coco, M. Leo, and C. Distante, "Facial expression recognition and histograms of oriented gradients: a comprehensive study," *SpringerPlus*, vol. 4, no. 1, p. 645, 2015.

[49] M. Niu, Y. Li, J. Tao, and S.-J. Wang, "Micro-expression recognition based on local two-order gradient pattern," in *IEEE Asian*

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

Conference on Affective Computing and Intelligent Interaction. IEEE, 2018, pp. 1–6.

[50] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[51] B. Sun, S. Cao, D. Li, J. He, and L. Yu, "Dynamic micro-expression recognition using knowledge distillation," *IEEE Transactions on Affective Computing*, 2020.

[52] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikäinen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 1, pp. 32–47, 2017.

[53] L.-w. Zhang, J. Li, S. Wang, X. Duan, W. Yan, H. Xie, and S. Huang, "Spatio-temporal fusion for macro-and micro-expression spotting in long video sequences," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020, pp. 245–252.

[54] M. Verburg and V. Menkovski, "Micro-expression detection in long videos using optical flow and recurrent neural networks," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–6.

[55] J. Li, C. Soladie, R. Seguier, S.-J. Wang, and M. H. Yap, "Spotting micro-expressions on long videos sequences," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–5.

[56] T.-K. Tran, Q.-N. Vo, X. Hong, and G. Zhao, "Dense prediction for micro-expression spotting based on deep sequence model," *Electronic Imaging*, vol. 2019, no. 8, pp. 401–1, 2019.

[57] K. X. Beh and K. M. Goh, "Micro-expression spotting using facial landmarks," in *IEEE International Colloquium on Signal Processing and Its Applications*. IEEE, 2019, pp. 192–197.

[58] S. Nag, A. K. Bhunia, A. Konwer, and P. P. Roy, "Facial micro-expression spotting and recognition using time contrasted feature with visual memory," in *IEEE International Conference on Acoustics, Speech and Signal Processing*. IEEE, 2019, pp. 2022–2026.

[59] Y. Han, B. Li, Y.-K. Lai, and Y.-J. Liu, "Cfd: a collaborative feature difference method for spontaneous micro-expression spotting," in *IEEE International Conference on Image Processing*. IEEE, 2018, pp. 1942–1946.

[60] C. A. Duque, O. Alata, R. Emonet, A.-C. Legrand, and H. Konik, "Micro-expression spotting using the riesz pyramid," in *IEEE Winter Conference on Applications of Computer Vision*. IEEE, 2018, pp. 66–74.

[61] 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 Antelligence*, vol. 31, no. 1, pp. 39–58, 2008.

[62] Y.-C. Lin, M.-C. Hu, W.-H. Cheng, Y.-H. Hsieh, and H.-M. Chen, "Human action recognition and retrieval using sole depth information," in *Proceedings of the 20th ACM international conference on Multimedia*, 2012, pp. 1053–1056.

[63] M.-C. Hu, C.-W. Chen, W.-H. Cheng, C.-H. Chang, J.-H. Lai, and J.-L. Wu, "Real-time human movement retrieval and assessment with kinect sensor," *IEEE transactions on Cybernetics*, vol. 45, no. 4, pp. 742–753, 2014.

[64] I.-C. Shen and W.-H. Cheng, "Gestalt rule feature points," *IEEE Transactions on Multimedia*, vol. 17, no. 4, pp. 526–537, 2015.

[65] A. R. Rivera, J. R. Castillo, and O. O. Chae, "Local directional number pattern for face analysis: Face and expression recognition," *IEEE transactions on Image Processing*, vol. 22, no. 5, pp. 1740–1752, 2012.

[66] T. Ojala, M. Pietikainen, and D. Harwood, "Performance evaluation of texture measures with classification based on kullback discrimination of distributions," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 1994, pp. 582–585.

[67] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, 2002.

[68] 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.

[69] Y. Zong, W. Zheng, X. Huang, J. Shi, Z. Cui, and G. Zhao, "Domain regeneration for cross-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 27, no. 5, pp. 2484–2498, 2018.

[70] Y. Zhang, H. Jiang, X. Li, B. Lu, K. M. Rabie, and A. U. Rehman, "A new framework combining local-region division and feature selection for micro-expressions recognition," *IEEE Access*, vol. 8, pp. 94 499–94 509, 2020.

[71] B. Xia, W. Wang, S. Wang, and E. Chen, "Learning from macro-expression: a micro-expression recognition framework," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2936–2944.

[72] R. Zhi, H. Xu, M. Wan, and T. Li, "Combining 3d convolutional neural networks with transfer learning by supervised pre-training for facial micro-expression recognition," *IEICE Transactions on Information and Systems*, vol. 102, no. 5, pp. 1054–1064, 2019.

[73] L. Zhou, Q. Mao, and L. Xue, "Cross-database micro-expression recognition: a style aggregated and attention transfer approach," in *IEEE International Conference on Multimedia and Expo Workshops*. IEEE, 2019, pp. 102–107.

[74] X. Jia, X. Ben, H. Yuan, K. Kpalma, and W. Meng, "Macro-to-micro transformation model for micro-expression recognition," *Journal of Computational Science*, vol. 25, pp. 289–297, 2018.

[75] C. Wang, M. Peng, T. Bi, and T. Chen, "Micro-attention for micro-expression recognition," *Neurocomputing*, vol. 410, pp. 354–362, 2020.

[76] M. Peng, Z. Wu, Z. Zhang, and T. Chen, "From macro to micro expression recognition: Deep learning on small datasets using transfer learning," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2018, pp. 657–661.

[77] X. Ben, X. Jia, R. Yan, X. Zhang, and W. Meng, "Learning effective binary descriptors for micro-expression recognition transferred by macro-information," *Pattern Recognition Letters*, vol. 107, pp. 50–58, 2018.

[78] L. Mao, N. Wang, L. Wang, and Y. Chen, "Classroom micro-expression recognition algorithms based on multi-feature fusion," *IEEE Access*, vol. 7, pp. 64 978–64 983, 2019.

[79] C. Guo, J. Liang, G. Zhan, Z. Liu, M. Pietikäinen, and L. Liu, "Extended local binary patterns for efficient and robust spontaneous facial micro-expression recognition," *IEEE Access*, vol. 7, pp. 174 517–174 530, 2019.

[80] C. Hu, D. Jiang, H. Zou, X. Zuo, and Y. Shu, "Multi-task micro-expression recognition combining deep and handcrafted features," in *IEEE International Conference on Pattern Recognition)*. IEEE, 2018, pp. 946–951.

[81] C. A. Duque, O. Alata, R. Emonet, H. Konik, and A.-C. Legrand, "Mean oriented riesz features for micro expression classification," *Pattern Recognition Letters*, vol. 135, pp. 382–389, 2020.

[82] B. Wu, T. Mei, W.-H. Cheng, and Y. Zhang, "Unfolding temporal dynamics: Predicting social media popularity using multi-scale temporal decomposition," 2016.

[83] B. Wu, W.-H. Cheng, Y. Zhang, and T. Mei, "Time matters: Multi-scale temporalization of social media popularity," in *Proceedings of the 24th ACM international conference on Multimedia*, 2016, pp. 1336–1344.

[84] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2005, pp. 886–893.

[85] B. K. Horn and B. G. Schunck, "Determining optical flow," in *Techniques and Applications of Image Understanding*, vol. 281. International Society for Optics and Photonics, 1981, pp. 319–331.

[86] Z. Xia, W. Peng, H.-Q. Khor, X. Feng, and G. Zhao, "Revealing the invisible with model and data shrinking for composite-database micro-expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 8590–8605, 2020.

[87] S.-T. Liong, Y. Gan, J. See, H.-Q. Khor, and Y.-C. Huang, "Shallow triple stream three-dimensional cnn (ststnet) for micro-expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–5.

[88] Y. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129–139, 2019.

[89] Y. Liu, H. Du, L. Zheng, and T. Gedeon, "A neural micro-expression recognizer," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–4.

[90] Y. Zhao and J. Xu, "A convolutional neural network for compound micro-expression recognition," *Sensors*, vol. 19, no. 24, p. 5553, 2019.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

[91] Y. Gan and S.-T. Liong, "Bi-directional vectors from apex in cnn for micro-expression recognition," in *IEEE International Conference on Image, Vision and Computing*. IEEE, 2018, pp. 168–172.

[92] Q. Li, S. Zhan, L. Xu, and C. Wu, "Facial micro-expression recognition based on the fusion of deep learning and enhanced optical flow," *Multimedia Tools and Applications*, vol. 78, no. 20, pp. 29 307–29 322, 2019.

[93] Q. Li, J. Yu, T. Kurihara, and S. Zhan, "Micro-expression analysis by fusing deep convolutional neural network and optical flow," in *IEEE International Conference on Control, Decision and Information Technologies*. IEEE, 2018, pp. 265–270.

[94] S. Happy and A. Routray, "Fuzzy histogram of optical flow orientations for micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 394–406, 2017.

[95] W. Zhong, X. Yu, L. Shi, and Z. Xie, "Facial micro-expression recognition based on local region of the key frame," in *International Symposium on Multispectral Image Processing and Pattern Recognition*, vol. 11430. International Society for Optics and Photonics, 2020, p. 114301L.

[96] Y. Gan, S.-T. Liong, W.-C. Yau, Y.-C. Huang, and L.-K. Tan, "Off-apexnet on micro-expression recognition system," *Signal Processing: Image Communication*, vol. 74, pp. 129–139, 2019.

[97] S.-T. Liong, J. See, K. Wong, and R. C.-W. Phan, "Less is more: Micro-expression recognition from video using apex frame," *Signal Processing: Image Communication*, vol. 62, pp. 82–92, 2018.

[98] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *IEEE conference on Computer Vision and Pattern Recognition*. IEEE, 2009, pp. 248–255.

[99] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: Towards real-time object detection with region proposal networks," in *Advances in neural information processing systems*, 2015, pp. 91–99.

[100] H.-C. Wang, Y.-C. Lai, W.-H. Cheng, C.-Y. Cheng, and K.-L. Hua, "Background extraction based on joint gaussian conditional random fields," *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 28, no. 11, pp. 3127–3140, 2017.

[101] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets and therapy*, vol. 8, 2015.

[102] F. S. Abousaleh, T. Lim, W.-H. Cheng, N.-H. Yu, M. A. Hossain, and M. F. Alhamid, "A novel comparative deep learning framework for facial age estimation," *EURASIP Journal on Image and Video Processing*, vol. 2016, no. 1, p. 47, 2016.

[103] C.-W. Hsieh, C.-Y. Chen, C.-L. Chou, H.-H. Shuai, J. Liu, and W.-H. Cheng, "Fashionon: Semantic-guided image-based virtual try-on with detailed human and clothing information," in *Proceedings of the 27th ACM International Conference on Multimedia*. ACM, 2019, pp. 275–283.

[104] I. Y. Susanto, T.-Y. Pan, C.-W. Chen, M.-C. Hu, and W.-H. Cheng, "Emotion recognition from galvanic skin response signal based on deep hybrid neural networks," in *International Conference on Multimedia Retrieval*. ACM, 2020, pp. 341–345.

[105] W.-H. Cheng, J. Liu, M. S. Kankanhalli, A. El Saddik, and B. Huet, "Ai+ multimedia make better life?" in *Proceedings of the 26th ACM international conference on Multimedia*, 2018, pp. 1455–1456.

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

[107] S. Ji, W. Xu, M. Yang, and K. Yu, "3d convolutional neural networks for human action recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 35, no. 1, pp. 221–231, 2012.

[108] Y.-C. Chen, D. S. Tan, W.-H. Cheng, and K.-L. Hua, "3d object completion via class-conditional generative adversarial network," in *International Conference on Multimedia Modeling*. Springer, 2019, pp. 54–66.

[109] K.-L. Hua, C.-H. Hsu, S. C. Hidayati, W.-H. Cheng, and Y.-J. Chen, "Computer-aided classification of lung nodules on computed tomography images via deep learning technique," *OncoTargets and therapy*, vol. 8, 2015.

[110] M. Peng, C. Wang, T. Chen, G. Liu, and X. Fu, "Dual temporal scale convolutional neural network for micro-expression recognition," *Frontiers in Psychology*, vol. 8, p. 1745, 2017.

[111] M. A. Takalkar and M. Xu, "Image based facial micro-expression recognition using deep learning on small datasets," in *IEEE International Conference on Digital Image Computing: Techniques and Applications*. IEEE, 2017, pp. 1–7.

[112] C. Wu and F. Guo, "Tsnn: Three-stream combining 2d and 3d convolutional neural network for micro-expression recognition," *IEEE Transactions on Electrical and Electronic Engineering*, 2020.

[113] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2018, pp. 667–674.

[114] S. Zhao, H. Tao, Y. Zhang, T. Xu, K. Zhang, Z. Hao, and E. Chen, "A two-stage 3d cnn based learning method for spontaneous micro-expression recognition," *Neurocomputing*, vol. 448, pp. 276–289, 2021.

[115] J. Li, Y. Wang, J. See, and W. Liu, "Micro-expression recognition based on 3d flow convolutional neural network," *Pattern Analysis and Applications*, vol. 22, no. 4, pp. 1331–1339, 2019.

[116] S. P. T. Reddy, S. T. Karri, S. R. Dubey, and S. Mukherjee, "Spontaneous facial micro-expression recognition using 3d spatiotemporal convolutional neural networks," in *IEEE International Joint Conference on Neural Networks*. IEEE, 2019, pp. 1–8.

[117] Z. Zhang, T. Chen, H. Meng, G. Liu, and X. Fu, "SMEConvNet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos," *IEEE Access*, vol. 6, pp. 71 143–71 151, 2018.

[118] T.-K. Tran, Q.-N. Vo, X. Hong, X. Li, and G. Zhao, "Micro-expression spotting: A new benchmark," *Neurocomputing*, vol. 443, pp. 356–368, 2021.

[119] S.-T. Liong, J. See, K. Wong, A. C. Le Ngo, Y.-H. Oh, and R. Phan, "Automatic apex frame spotting in micro-expression database," in *IEEE Asian Conference on Pattern Recognition*. IEEE, 2015, pp. 665–669.

[120] S. Li and W. Deng, "Deep facial expression recognition: A survey," *IEEE Transactions on Affective Computing*, 2020.

[121] R. Weber, J. Li, C. Soladie, and R. Seguier, "A survey on databases of facial macro-expression and micro-expression," in *International Joint Conference on Computer Vision, Imaging and Computer Graphics*. Springer, 2018, pp. 298–325.

[122] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, "A survey on deep transfer learning," in *International Conference on Artificial Neural Networks*. Springer, 2018, pp. 270–279.

[123] 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.

[124] 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 *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, 2010, pp. 94–101.

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

[126] N. Aifanti, C. Papachristou, and A. Delopoulos, "The mug facial expression database," in *11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10*. IEEE, 2010, pp. 1–4.

[127] M. F. Valstar, E. Sánchez-Lozano, J. F. Cohn, L. A. Jeni, J. M. Girard, Z. Zhang, L. Yin, and M. Pantic, "Fera 2017-addressing head pose in the third facial expression recognition and analysis challenge," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2017, pp. 839–847.

[128] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros, "Unpaired image-to-image translation using cycle-consistent adversarial networks," in *IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2223–2232.

[129] B. Allaert, I. M. Bilasco, and C. Djeraba, "Micro and macro facial expression recognition using advanced local motion patterns," *IEEE Transactions on Affective Computing*, 2019.

[130] S. S. Pawar, M. Moh, and T.-S. Moh, "Micro-expression recognition using motion magnification and spatiotemporal texture map," in *International Conference on Ubiquitous Information Management and Communication*. Springer, 2019, pp. 351–369.

[131] R. Asmara, P. Choirina, C. Rahmad, A. Setiawan, F. Rahutomo, R. Yusron, and U. Rosiani, "Study of drmf and asm facial landmark point for micro expression recognition using klt tracking point feature," in *Journal of Physics: Conference Series*, vol. 1402, no. 7. IOP Publishing, 2019, p. 077039.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

[132] H. Lu, K. Kpalma, and J. Ronsin, "Motion descriptors for micro-expression recognition," *Signal Processing: Image Communication*, vol. 67, pp. 108–117, 2018.

[133] Y. Zong, X. Huang, W. Zheng, Z. Cui, and G. Zhao, "Learning from hierarchical spatiotemporal descriptors for micro-expression recognition," *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3160–3172, 2018.

[134] L. Zhou, Q. Mao, and L. Xue, "Dual-inception network for cross-database micro-expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–5.

[135] L. Wang, H. Xiao, S. Luo, J. Zhang, and X. Liu, "A weighted feature extraction method based on temporal accumulation of optical flow for micro-expression recognition," *Signal Processing: Image Communication*, vol. 78, pp. 246–253, 2019.

[136] M. Verma, S. K. Vipparthi, G. Singh, and S. Murala, "Learnet: Dynamic imaging network for micro expression recognition," *IEEE Transactions on Image Processing*, vol. 29, pp. 1618–1627, 2019.

[137] C. Lin, F. Long, J. Huang, and J. Li, "Micro-expression recognition based on spatiotemporal gabor filters," in *IEEE International Conference on Information Science and Technology*. IEEE, 2018, pp. 487–491.

[138] M. A. Takalkar, S. Thuseethan, S. Rajasegarar, Z. Chaczko, M. Xu, and J. Yearwood, "LGAttNet: Automatic micro-expression detection using dual-stream local and global attentions," *Knowledge-Based Systems*, vol. 212, p. 106566, 2021.

[139] J. Wang, X. Pan, X. Li, G. Wei, and Y. Zhou, "Single trunk multi-scale network for micro-expression recognition," *Graphics and Visual Computing*, p. 200026, 2021.

[140] L. Lei, J. Li, T. Chen, and S. Li, "A novel graph-tcn with a graph structured representation for micro-expression recognition," in *Proceedings of the 28th ACM International Conference on Multimedia*. ACM, 2020, pp. 2237–2245.

[141] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Transactions on Multimedia*, vol. 22, no. 3, pp. 626–640, 2019.

[142] Z. Xia, H. Liang, X. Hong, and X. Feng, "Cross-database micro-expression recognition with deep convolutional networks," in *ACM International Conference on Biometric Engineering and Applications*. ACM, 2019, pp. 56–60.

[143] L. Lu, L. Tavabi, and M. Soleymani, "Self-supervised learning for facial action unit recognition through temporal consistency," in *British Machine Vision Conference*. BMVA, 2020.

[144] C. Wang, J. Zeng, S. Shan, and X. Chen, "Multi-task learning of emotion recognition and facial action unit detection with adaptively weights sharing network," in *IEEE International Conference on Image Processing*. IEEE, 2019, pp. 56–60.

[145] C. Ma, L. Chen, and J. Yong, "Au r-cnn: Encoding expert prior knowledge into r-cnn for action unit detection," *Neurocomputing*, vol. 355, pp. 35–47, 2019.

[146] I. O. Ertugrul, J. F. Cohn, L. A. Jeni, Z. Zhang, L. Yin, and Q. Ji, "Cross-domain au detection: Domains, learning approaches, and measures," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 2019, pp. 1–8.

[147] W. Li, F. Abtahi, and Z. Zhu, "Action unit detection with region adaptation, multi-labeling learning and optimal temporal fusing," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2017, pp. 1841–1850.

[148] W. Li, F. Abtahi, Z. Zhu, and L. Yin, "Eac-net: A region-based deep enhancing and cropping approach for facial action unit detection," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2017, pp. 103–110.

[149] Y. Fan, J. C. Lam, and V. O. K. Li, "Facial action unit intensity estimation via semantic correspondence learning with dynamic graph convolution." in *AAAI Conference on Artificial Intelligence*. AAAI, 2020, pp. 12 701–12 708.

[150] Y. Li, J. Zeng, and S. Shan, "Learning representations for facial actions from unlabeled videos," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2020.

[151] R. Zhi, M. Liu, and D. Zhang, "A comprehensive survey on automatic facial action unit analysis," *The Visual Computer*, vol. 36, no. 5, pp. 1067–1093, 2020.

[152] B. Martinez, M. F. Valstar, B. Jiang, and M. Pantic, "Automatic analysis of facial actions: A survey," *IEEE Transactions on Affective Computing*, vol. 10, no. 3, pp. 325–347, 2019.

[153] Z. Liu, J. Dong, C. Zhang, L. Wang, and J. Dang, "Relation modeling with graph convolutional networks for facial action unit detection," in *International Conference on Multimedia Modeling*. Springer, 2020, pp. 489–501.

[154] G. Li, X. Zhu, Y. Zeng, Q. Wang, and L. Lin, "Semantic relationships guided representation learning for facial action unit recognition," in *AAAI Conference on Artificial Intelligence*, vol. 33. AAAI, 2019, pp. 8594–8601.

[155] Z.-M. Chen, X.-S. Wei, P. Wang, and Y. Guo, "Multi-label image recognition with graph convolutional networks," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2019, pp. 5177–5186.

[156] W.-S. Chu, F. De la Torre, and J. F. Cohn, "Learning spatial and temporal cues for multi-label facial action unit detection," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2017, pp. 25–32.

[157] K. Zhao, W.-S. Chu, F. De la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit and holistic expression recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3931–3946, 2016.

[158] Y.-J. Liu, J.-K. Zhang, W.-J. Yan, S.-J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, 2015.

[159] L. Lo, H.-X. Xie, H.-H. Shuai, and W.-H. Cheng, "Mer-gcn: Micro-expression recognition based on relation modeling with graph convolutional networks," in *IEEE Conference on Multimedia Information Processing and Retrieval*. IEEE, 2020, pp. 79–84.

[160] Q. Li, J. Yu, T. Kurihara, H. Zhang, and S. Zhan, "Deep convolutional neural network with optical flow for facial micro-expression recognition," *Journal of Circuits, Systems and Computers*, vol. 29, no. 01, p. 2050006, 2020.

[161] W. Merghani and M. H. Yap, "Adaptive mask for region-based facial micro-expression recognition," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2020, pp. 428–433.

[162] M. A. Takalkar, M. Xu, and Z. Chaczko, "Manifold feature integration for micro-expression recognition," *Multimedia Systems*, vol. 26, no. 5, pp. 535–551, 2020.

[163] C. H. Lim and K. M. Goh, "Fuzzy qualitative approach for micro-expression recognition," in *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference*. IEEE, 2017, pp. 1669–1674.

[164] T.-Y. Pan, Y.-Z. Dai, M.-C. Hu, and W.-H. Cheng, "Furniture style compatibility recommendation with cross-class triplet loss," *Multimedia Tools and Applications*, vol. 78, no. 3, pp. 2645–2665, 2019.

[165] A. J. Smola and B. Schölkopf, "A tutorial on support vector regression," *Statistics and computing*, vol. 14, no. 3, pp. 199–222, 2004.

[166] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf, "Deepface: Closing the gap to human-level performance in face verification," in *IEEE International Conference on Computer Vision and Pattern Recognition*. IEEE, 2014, pp. 1701–1708.

[167] S. Lalitha and K. Thyagharajan, "Micro-facial expression recognition based on deep-rooted learning algorithm," *International Journal of Computational Intelligence Systems*, vol. 12, pp. 903–913, 2019.

[168] T.-Y. Lin, P. Goyal, R. Girshick, K. He, and P. Dollár, "Focal loss for dense object detection," in *IEEE International Conference on Computer Vision*. IEEE, 2017, pp. 2980–2988.

[169] Z. Lai, R. Chen, J. Jia, and Y. Qian, "Real-time micro-expression recognition based on resnet and atrous convolutions," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–12, 2020.

[170] W. Xie, L. Shen, and J. Duan, "Adaptive weighting of handcrafted feature losses for facial expression recognition," *IEEE Transactions on Cybernetics*, 2019.

[171] T. Zhang, Y. Zong, W. Zheng, C. P. Chen, X. Hong, C. Tang, Z. Cui, and G. Zhao, "Cross-database micro-expression recognition: A benchmark," *IEEE Transactions on Knowledge and Data Engineering*, 2020.

[172] X. Zhang, T. Xu, W. Sun, and A. Song, "Multiple source domain adaptation in micro-expression recognition," *Journal of Ambient Intelligence and Humanized Computing*, pp. 1–16, 2020.

[173] L. Li, X. Zhou, Y. Zong, W. Zheng, X. Chen, J. Shi, and P. Song, "Unsupervised cross-database micro-expression recognition us-

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/

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.2022.3143100, IEEE Transactions on Affective Computing

ing target-adapted least-squares regression," *IEICE Transactions on Information and Systems*, vol. 102, no. 7, pp. 1417–1421, 2019.
[174] G. Wilson and D. J. Cook, "A survey of unsupervised deep domain adaptation," *ACM Transactions on Intelligent Systems and Technology*, vol. 11, no. 5, pp. 1–46, 2020.
[175] L. Zhang, "Transfer adaptation learning: A decade survey," *arXiv preprint arXiv:1903.04687*, 2019.
[176] A. J. R. Kumar, R. Theagarajan, O. Peraza, and B. Bhanu, "Classification of facial micro-expressions using motion magnified emotion avatar images." in *IEEE International Conference on Computer Vision and Pattern Recognition Workshops*. IEEE, 2019, pp. 12–20.
[177] M. Humphreys, "Reflections on the ethics of social experimentation," *Journal of Globalization and Development*, vol. 6, no. 1, pp. 87–112, 2015.
[178] L. A. Zadeh, "Fuzzy sets," *Information and Control*, vol. 8, no. 3, pp. 338–353, 1965.
[179] M. Chen, H. T. Ma, J. Li, and H. Wang, "Emotion recognition using fixed length micro-expressions sequence and weighting method," in *IEEE International Conference on Real-time Computing and Robotics*. IEEE, 2016, pp. 427–430.
[180] H. A. Elfenbein and N. Ambady, "Is there an in-group advantage in emotion recognition?" *Psychological Bulletin*, vol. 128, pp. 243–9, 2002.
[181] S. G. Young and J. P. Wilson, "A minimal ingroup advantage in emotion identification confidence," *Cognition and Emotion*, vol. 32, no. 1, pp. 192–199, 2018.
[182] E. R. Tuminello and D. Davidson, "What the face and body reveal: In-group emotion effects and stereotyping of emotion in african american and european american children," *Journal of Experimental Child Psychology*, vol. 110, no. 2, pp. 258–274, 2011.
[183] P. Thibault, P. Bourgeois, and U. Hess, "The effect of group-identification on emotion recognition: The case of cats and basketball players," *Journal of Experimental Social Psychology*, vol. 42, no. 5, pp. 676–683, 2006.
[184] S. G. Young and K. Hugenberg, "Mere social categorization modulates identification of facial expressions of emotion." *Journal of Personality and Social Psychology*, vol. 99, no. 6, p. 964, 2010.
[185] S. Huang and S. Han, "Shared beliefs enhance shared feelings: religious/irreligious identifications modulate empathic neural responses," *Social Neuroscience*, vol. 9, no. 6, pp. 639–649, 2014.
[186] C. Prado, D. Mellor, L. K. Byrne, C. Wilson, X. Xu, and H. Liu, "Facial emotion recognition: a cross-cultural comparison of chinese, chinese living in australia, and anglo-australians," *Motivation and Emotion*, vol. 38, no. 3, pp. 420–428, 2014.
[187] Y. Xie, C. Zhong, F. Zhang, and Q. Wu, "The ingroup disadvantage in the recognition of micro-expressions," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–5.
[188] M. Fölster, U. Hess, and K. Werheid, "Facial age affects emotional expression decoding," *Frontiers in Psychology*, vol. 5, p. 30, 2014.
[189] S. Chen, J. Wang, Y. Chen, Z. Shi, X. Geng, and Y. Rui, "Label distribution learning on auxiliary label space graphs for facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 13 984–13 993.
[190] K. Wang, X. Peng, J. Yang, S. Lu, and Y. Qiao, "Suppressing uncertainties for large-scale facial expression recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2020, pp. 6897–6906.
[191] M. K. Mandal, H. S. Asthana, and R. Pandey, "Asymmetry in emotional face: Its role in intensity of expression," *The Journal of Psychology*, vol. 129, no. 2, pp. 235–241, 1995.
[192] W. G. Dopson, B. E. Beckwith, D. M. Tucker, and P. C. Bullard-Bates, "Asymmetry of facial expression in spontaneous emotion," *Cortex*, vol. 20, no. 2, pp. 243–251, 1984.
[193] H. Chen, X. Liu, X. Li, H. Shi, and G. Zhao, "Analyze spontaneous gestures for emotional stress state recognition: A micro-gesture dataset and analysis with deep learning," in *IEEE International Conference on Automatic Face and Gesture Recognition*. IEEE, 2019, pp. 1–8.

Hong-Xia Xie received the B.S. degree in Internet of Things from the Zhengzhou University of Aeronautics in 2016 and received the M.S. degree in communication and information systems, Fujian Normal University, China, in 2019. She is now pursuing a Ph.D. degree in Institute of of Electronics, National Yang Ming Chiao Tung Univresity, Taiwan. Her research interests include micro-expression recognition, emotion recognition and deep learning.

Ling Lo received the B.S. degree from Department of Electronics Engineering, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan, R.O.C., in 2019, and now she is pursuing a Ph.D degree in Institute of Electronics, NYCU. Her current research interests include deep learning and computer vision. Recently her work focuses specifically on facial and micro-expression recognition.

Hong-Han Shuai received the B.S. degree from the Department of Electrical Engineering, National Taiwan University (NTU), Taipei, Taiwan, R.O.C., in 2007, the M.S. degree in computer science from NTU in 2009, and the Ph.D. degree from Graduate Institute of Communication Engineering, NTU, in 2015. He is now an assistant professor in NYCU. His research interests are in the area of multimedia processing, machine learning, social network analysis, and data mining. His works have appeared in top-tier conferences such as MM, CVPR, AAAI, KDD, WWW, ICDM, CIKM and VLDB, and top-tier journals such as TKDE, TMM and JIOT. Moreover, he has served as the PC member for international conferences including MM, AAAI, IJCAI, WWW, and the invited reviewer for journals including TKDE, TMM, JVCI and JIOT.

Wen-Huang Cheng is Distinguished Professor with the Institute of Electronics, National Yang Ming Chiao Tung University (NYCU), Hsinchu, Taiwan. He is also Jointly Appointed Professor with the Artificial Intelligence and Data Science Program, National Chung Hsing University (NCHU), Taichung, Taiwan. Before joining NYCU, he led the Multimedia Computing Research Group at the Research Center for Information Technology Innovation (CITI), Academia Sinica, Taipei, Taiwan, from 2010 to 2018. His current research interests include multimedia, artificial intelligence, computer vision, and machine learning. He has actively participated in international events and played important leading roles in prestigious journals and conferences and professional organizations, like Associate Editor for IEEE Transactions on Multimedia, General co-chair for IEEE ICME (2022) and ACM ICMR (2021), Chair for IEEE MSA technical committee, governing board member for IAPR. He has received numerous research and service awards, including the Best Paper Award of 2021 IEEE ICME, the Outstanding Associate Editor Award of IEEE Transactions on Multimedia, the 2018 MSRA Collaborative Research Award, the 2017 Ta-Yu Wu Memorial Award from Taiwan's Ministry of Science and Technology (the highest national research honor for young Taiwanese researchers under age 42), the Top 10% Paper Award from the 2015 IEEE MMSP, and the K. T. Li Young Researcher Award from the ACM Taipei/Taiwan Chapter in 2014. He is IEEE Distinguished Lecturer, ACM Distinguished Member, and IET Fellow.

This work is licensed under a Creative Commons Attribution 4.0 License. For more information, see https://creativecommons.org/licenses/by/4.0/