Video-based Facial Micro-Expression Analysis: A Survey of Datasets, Features and Algorithms

Xianye Ben, *Member, IEEE*, Yi Ren, Junping Zhang, *Member, IEEE*, Su-Jing Wang, *Senior Member, IEEE*, Kidiyo Kpalma, Weixiao Meng, *Senior Member, IEEE*, Yong-Jin Liu, *Senior Member, IEEE*

Abstract—Unlike the conventional facial expressions, micro-expressions are involuntary and transient facial expressions capable of revealing the genuine emotions that people attempt to hide. Therefore, they can provide important information in a broad range of applications such as lie detection, criminal detection, etc. Since micro-expressions are transient and of low intensity, however, their detection and recognition is difficult and relies heavily on expert experiences. Due to its intrinsic particularity and complexity, video-based micro-expression analysis is attractive but challenging, and has recently become an active area of research. Although there have been numerous developments in this area, thus far there has been no comprehensive survey that provides researchers with a systematic overview of these developments with a unified evaluation. Accordingly, in this survey paper, we first highlight the key differences between macro- and micro-expressions, then use these differences to guide our research survey of video-based micro-expression analysis in a cascaded structure, encompassing the neuropsychological basis, datasets, features, spotting algorithms, recognition algorithms, applications and evaluation of state-of-the-art approaches. For each aspect, the basic techniques, advanced developments and major challenges are addressed and discussed. Furthermore, after considering the limitations of existing micro-expression datasets, we present and release a new dataset — called *micro-and-macro expression warehouse* (MMEW) — containing more video samples and more labeled emotion types. We then perform a unified comparison of representative methods on CAS(ME)^2 for spotting, and on MMEW and SAMM for recognition, respectively. Finally, some potential future research directions are explored and outlined.

Index Terms—Micro-expression analysis, survey, spotting, recognition, facial features, datasets

✦

1 INTRODUCTION

E MOTIONS are an inherent part of human life, and appear voluntarily or involuntarily through facial expressions when people communicate with each other face-to-face. As a typical form of nonverbal communication, facial expressions play an important role in the analysis of human emotion [1], and have thus been widely studied in various domains (e.g., [2], [3]).

Broadly speaking, there are two classes of facial expressions: macro- and micro-expressions (Figure 1). The major difference between these two classes lies in both their duration and intensity. Macro-expressions are voluntary, usually last for between 0.5 and 4 seconds [4], and are

| Happiness | Surprise | Anger | Disgust | Fear | Sadness |

(a) Macro-expressions

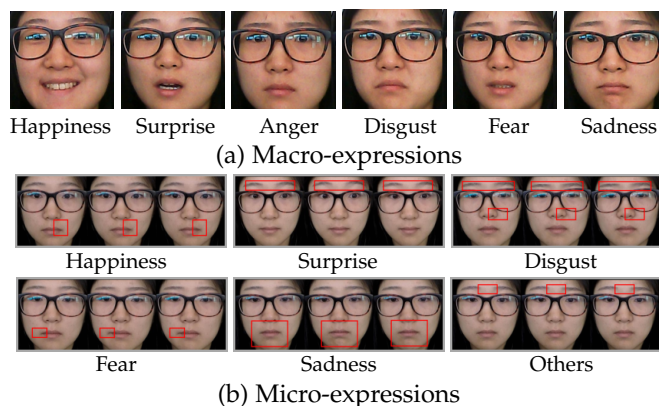| Happiness | Surprise | Disgust |
| Fear | Sadness | Others |

(b) Micro-expressions

Fig. 1: Six macro-expressions and six micro-expressions, sampled from the same person, in the MMEW dataset (see Section 3). While macro-expressions can be analyzed based on a single image, micro-expressions need to be analyzed across an image sequence due to their low intensity. For micro-expressions, the subtle changes outlined in red boxes are explained in the supplemental material.

- X. Ben and Y. Ren are with the School of Information Science and Engineering, Shandong University, Qingdao 266237, China. Email: benxianye@gmail.com, yiren@mail.sdu.edu.cn.
- J. Zhang is with the Shanghai Key Laboratory of Intelligent Information Processing, School of Computer Science, Fudan University, Shanghai 200433, China. Email: jpzhang@fudan.edu.cn.
- S.J. Wang is with the State Key Laboratory of Brain and Cognitive Science, Institute of Psychology, Chinese Academy of Sciences, Beijing, 100101, China. Email: wangsujing@psych.ac.cn.
- Kidiyo Kpalma is with IETR CNRS UMR 6164, Institut National des Sciences Appliquées de Rennes, Rennes 35708, France. Email: Kidiyo.Kpalma@insa-rennes.fr.
- W. Meng is with the School of Electronics and Information Engineering, Harbin Institute of Technology, Harbin 150080, China. E-mail: wxmeng@hit.edu.cn.
- Y.J. Liu is with MOE-Key Laboratory of Pervasive Computing, BNRist, the Department of Computer Science and Technology, Tsinghua University, Beijing, 100084, China. E-mail: liuyongjin@tsinghua.edu.cn.

Yong-Jin Liu and Su-Jing Wang are corresponding authors.

made using underlying facial movements that cover a large facial area [3]; thus, they can be clearly distinguished from noise such as eye blinks. By contrast, micro-expressions are involuntary, rapid and local expressions [5], [6], the typical duration of which is between 0.065 and 0.5 seconds [7]. Although people may intentionally conceal or restrain their real emotions by disguising their macro-expressions, micro-expressions are uncontrollable and can thus reveal the

genuine emotions that humans try to conceal [5], [6], [8], [9].

Due to their inherent properties (short duration, involuntariness and low intensity), micro-expressions are very difficult to identify with the naked eye; only experts who have been extensively trained can distinguish micro-expressions. Nevertheless, even after intense training, humans can recognize only 47% of micro-expressions on average [10]. Moreover, human analysis of micro-expressions is time-consuming, expensive and error-prone. Therefore, it is highly desirable to develop automatic systems for micro-expressions analysis based on computer vision and pattern analysis techniques [11].

Note that, in this paper, the concept of micro-expression analysis involves two aspects: namely, spotting and recognition. Micro-expression spotting involves identifying whether a given video clip contains a micro-expression, and if such an expression is found, identifying the onset (starting time), apex (the time with the highest intensity of expression) and offset (ending time) frames. Micro-expression recognition involves classifying a micro-expression into a set of predefined emotion types, e.g., happiness, surprise, sadness, disgust or anger, etc.

1.1 Challenges and differences from conventional techniques

There have been a large number of various techniques proposed for conventional macro-expression analysis (see e.g. [3] and reference therein). However, it is non-trivial to adapt these existing techniques to micro-expression analysis. Below, we list the three main technical challenges when compared to conventional macro-expression analysis.

(1) Challenges in collecting datasets. It is very difficult to elicit proper micro-expressions from participants in a controlled environment. Moreover, it is also difficult to correctly label these elicited micro-expressions as the ground truth, even for experts. Therefore, useful micro-expression datasets are scarce. In the few datasets available, as summarized in Section 3, the number of samples is much smaller than the case for conventional macro-expression datasets, which presents a key challenge for designing or training micro-expression analysis algorithms.

(2) Challenges in designing spotting algorithms. Macro-expression can usually be accomplished using a single face image. However, due to their low intensity, it is almost impossible to detect micro-expressions in a single image; instead, an image sequence is required. Furthermore, different from macro-expression detection, spotting is a novel problem in micro-expression analysis. The techniques of micro-expression spotting are summarized in Section 5.

(3) Challenges in designing recognition algorithms. Like micro-expression spotting, micro-expression recognition requires an image sequence. Usually, the image/video features used in macro-expression analysis can also be used in micro-expression analysis: however, special treatment is required for the latter. Due to their short duration and low intensity, the signals of these features are very weak, while indiscriminately amplifying these signals will amplify noises from head movement, illumination, optical flow estimation, etc. Section 4 summarizes facial features for micro-expression and Section 6 summarizes the techniques of micro-expression recognition.

1.2 Contributions

Over the past decade, research into micro-expression analysis has been blossoming in terms of datasets, spotting and recognition techniques. However, the large part of micro-expression research remains scattered, and very few systematic surveys exist. One recent survey work [12] focuses on the pipeline of a micro-expression recognition system from an engineering perspective. A second survey [13] provides a collection of results from classic methods, but it lacks a consistent experimental analysis. Different from [12] and [13], this paper presents a comprehensive survey that makes the following contributions:

- We provide a comprehensive overview focusing on computing methodologies related to micro-expression spotting and recognition, as well as the common image and video features that can be used to build an appropriate automatic system. This survey also includes a detailed summary of up-to-date micro-expression datasets; based on this summary, we perform a unified comparison of representative micro-expression spotting and recognition methods.
- Based on our summary of the limitations of existing micro-expression datasets, we present and release a new dataset, called micro-and-macro expression warehouse (MMEW), which contains both macro- and micro-expressions sampled from the same subjects. This new characteristic can inspire promising future research to explore the relationship between macro- and micro-expressions of the same subject.

The remainder of this paper is organized as follows. Section 2 highlights the neuropsychological difference between macro- and micro-expressions. Section 3 summarizes existing datasets. The image and video features useful for micro-expression analysis are collected and categorized in Section 4. Representative algorithms are summarized in Sections 5 (for micro-expression spotting) and 6 (for micro-expression recognition). Potential applications of micro-expression analysis are summarized in Section 7. Section 8 presents a detailed comparison and recommendation of different methods. Some remaining challenges and future research directions are summarized in Section 9. Finally Section 10 presents our concluding remarks.

2 DIFFERENCES BETWEEN MACRO- AND MICRO-EXPRESSIONS

Facial expressions are the results of the movement of facial skin and connective tissue. Facial muscles, which control these movements, are activated by facial nerve nuclei, which are in turn controlled by cortical and subcortical upper motor neuron circuits. One neuropsychological study of facial expression [14] showed that there are two distinct neural pathways (located in different brain areas) for mediating facial behavior. The cortical circuit is located in the cortical motor strip and is primarily responsible for *posed* facial expressions (i.e., voluntary facial actions). Moreover, the subcortical circuit, which is located in the subcortical areas of the brain, is primarily responsible for *spontaneous* facial expressions (i.e., involuntary emotion). When people attempt to conceal or restrain their expressions in an intensely
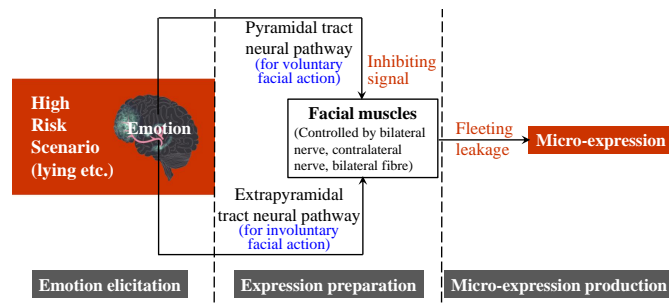
Fig. 2: Two distinct neural pathways exist for conveying facial behavior: namely, the pyramidal and extrapyramidal tract neural pathways. The former is responsible for macro-expressions with voluntary facial actions, while the latter is responsible for spontaneous facial expressions. In high-risk scenarios, such as lying, both pathways are activated and engage in a back-and-forth battle, resulting in the fleeting leakage of genuine emotions in the form of micro-expressions.

emotional situation, both systems are likely to be activated, resulting in the fleeting leakage of genuine emotions in the form of micro-expressions [15] (Figure 2). Throughout this paper, we will focus on spontaneous micro-expressions.

Micro-expressions are localized facial deformations caused by the involuntary contraction of facial muscles [16]. In comparison, macro-expressions involve more muscle in a larger facial area and the intensity of muscle motion is also relatively stronger. Compared to macro-expressions, the intrinsic differences in terms of neuroanatomical mechanism is that micro-expressions have very short duration, slight variation and fewer action areas on the external facial features [17], [18]. This difference can also be concluded from the mechanism of concealment [5]: when people try to conceal their feelings, their true emotion can quickly 'leak out' and may be manifested as micro-expressions.

Other studies have also shown that the motion of muscles in the upper face (forehead and upper eyelid) during fake expressions are controlled by the precentral gyrus, which is dominated by the bilateral nerve, while the motion of muscle in the lower face is controlled by the contralateral nerve [6], [9], [19]. The spontaneous expressions are controlled by the subcortical structure, the muscular movement of which is controlled by the bilateral fibre. This also provides evidence for the restraint hypothesis [20], which states that random expressions controlled by the pyramidal tract neural pathway undergo a back-and-forth battle with the spontaneous expressions controlled by the extrapyramidal tract; during this process, the extrapyramidal tract is dominated by bilateral fibers. Furthermore, if activity in the extrapyramidal tract is reflected on the face, it controls the muscles of the upper face, leading to spontaneous expressions [6]. This also leads to the conclusion that when micro-expressions occur, they can happen on different parts of the face [21].

The hypothesis that feedback from the upper and lower facial regions has different effects on micro-expression recognition was proposed in [22]. Three supporting studies were subsequently presented to highlight the three roles of facial feedback in judging the subtle movements of micro-

TABLE 1: Main differences between macro- and micro-expressions.

Difference	Micro-expression	Macro-expression
Noticeability	Easy to ignore	Easily noticed
Time interval	Short duration (0.065-0.5 seconds)	Long duration (0.5-4 seconds)
Motion intensity	Slight variation	Large variation
Subjectivity	Involuntary (uncontrollable)	Voluntary (under control)
Action areas	Fewer	Almost all areas

expressions. First, when feedback in the upper face is enhanced by a restricting gel, micro-expressions with a duration of 450 ms can be more easily detected. Second, when lower face feedback is enhanced, the accuracy of sensing micro-expressions (duration conditions of 50, 150, 333, and 450 ms) is reduced. Third, blocking lower face feedback improves the recognition accuracy of micro-expressions.

Characteristics of micro-expressions. Micro-expressions can express seven universal emotions: disgust, anger, fear, sadness, happiness, surprise and contempt [23]. Ekman suggests that there are certain facial muscles that cannot be consciously controlled, which he refers to as reliable muscles; these muscles serve as sound indicators of the occurrence of related emotions [24]. Studies have also shown that micro-expressions may contain all or only part of the muscle movements that make up common expressions [24], [25], [26]. Therefore, compared to macro-expressions, micro-expressions are facial expressions with short duration and characterized by greater facial muscle movement inhibition [24], [25], which can reflect a person's true emotions and are more difficult to control [27]. Table 1 summarizes the major differences between macro- and micro-expressions.

Short duration. The short duration is considered to be the most important characteristic of micro-expressions. Most psychologists now agree that micro-expressions do not last for more than half a second. Yan et al. [7] performed an elegant analysis that employed distribution curves of total duration and onset duration for the purposes of micro-expression analysis. These authors revealed that in a micro-expression, the start-up time (time from the onset frame to the apex frame) is usually within 0.26 seconds. Furthermore, the intensity of muscle motion in the micro-expression is very weak and the expression itself is uncontrollable due to the psychological inhibition of human instinctive response.

Dynamic features. Neuropsychology shows that left and right sides of the face differ from each other in terms of expressing emotions [6]: the left side expresses emotions that are more intense. Studies on facial asymmetry have also found that the right side expresses social context cues more conspicuously, while the left side expresses more personal feelings [6]. These studies provide further evidence for the distinction between fake expressions and natural expressions, and thus indirectly explain the dynamic features of micro-expressions.

3 DATASETS OF MICRO-EXPRESSIONS

The development of micro-expression analysis techniques has been largely dependent on well-established datasets

with correctly labelled ground truth. Due to their intrinsic characteristics, such as involuntariness, short duration and slight variation, eliciting micro-expressions in a controlled environment is very difficult. Thus far, a few micro-expression datasets have been developed. However, as summarized in this section, most of these still have various deficiencies as regards their elicitation paradigms, labelling methods or small data size. Therefore, at the end of this section, we present and release a new dataset for micro-expression recognition.

Some early studies elicited micro-expressions by constructing high-stakes situations; for example, asking people to lie by concealing negative affect aroused by an unpleasant film and simulating pleasant feelings [34], or telling lies about a mock theft in a crime scenario [35]. However, micro-expressions elicited in this way are often contaminated by other non-emotional facial movements, such as conversational behavior.

Since 2011, nine representative micro-expression datasets have been developed: namely, USF-HD [36], Polikovsky's dataset [23], York DDT [37], MEVIEW [28], SMIC [30], CASME [31], CASME II [32], SAMM [29] and CAS(ME)^2 [33]. Both USF-HD and Polikovsky's dataset contain posed micro-expressions:

- in USF-HD, participants were asked to perform both macro- and micro-expressions, and
- in Polikovsky's dataset, participants were asked to simulate micro-expression motion.

It should be noted that these posed micro-expressions are different from spontaneous ones. Moreover, York DDT is made up of spontaneous micro-expressions with high ecological validity. Nevertheless, similar to lie detection [34], [35], the data in York DDT was mixed with non-emotional facial movement resulting from talking. Furthermore, all of three datasets (USF-HD, Polikovsky's dataset and York DDT) are not publicly available.

The paradigm of elicitation by telling lies has two major drawbacks:

- micro-expressions are often contaminated with irrelevant (i.e., talking) facial movements, and
- the types of elicited micro-expressions are seriously restricted (e.g., the happiness type can never be elicited).

By contrast, it is well recognized that watching an emotional video while maintaining a neutral expression (i.e. suppressing emotions) is an effective method for eliciting micro-expressions.

Five datasets — SMIC, CASME, CASME II, SAMM and CAS(ME)^2 — used this elicitation paradigm. MEVIEW used another quite different elicitation paradigm: namely, constructing a high-stakes situation by making use of poker games or TV interviews with difficult questions.

In the below, we first review MEVIEW, summarizing its advantages and disadvantages, and then move on to focus on the mainstream datasets (SMIC, CASME, CASME II, SAMM and CAS(ME)^2). All six of these datasets, which are publicly available, are summarized in Table 2; see also Figure 3 for some snapshots. For comparison, Table 2 also

includes our newly released dataset MMEW, which is presented in more detail at the end of this section.

MEVIEW [28] consists of realistic video clips (i.e., shots in a non-lab controlled environment) from two scenarios: (1) some key moments in poker games, and (2) a person being asked a difficult question in a TV interview. Both scenarios are notable for their high stress factor. For example, in poker games, players try to conceal or fake their true emotions, and the key moments in the videos show the detail of a player's face while the cards are being uncovered: at these moments, micro-expressions are most likely to appear. The MEVIEW dataset contains 40 micro-expression video clips at 25 fps with a resolution of 1280×720. The average length of the video clips in the dataset is 3 seconds and the camera shot is often switched. The emotion types in MEVIEW are divided into seven classes: happiness, contempt, disgust, surprise, fear, anger and unclear emotions.

The advantage of the MEVIEW dataset is that its scenarios are real, which will benefit the training or testing of micro-expression analysis algorithms. The disadvantages include the following: (1) in these real-scene videos, the participants were seldom shot from the frontal pose, meaning that the number of valid samples is quite small, and (2) the number of participants in the dataset is only 16, which is small.

We will now review SMIC, SAMM, CASME, CASME II and CAS(ME)^2. All the image sequences in these datasets are shots in controlled laboratory environments.

The SMIC dataset [30] provides three data subsets with different types of recording cameras: HS (standing for high-speed camera), VIS (for normal visual camera) and NIR (for near-infrared camera). Since micro-expressions have a short duration and low intensity, a higher spatial and temporal resolution may help to capture more details. Therefore, the HS subset was collected with a high-speed camera with a frame rate of 100 fps. HS can be used to study the characteristics of the rapid changes of the micro-expressions. Moreover, the VIS and NIR subsets are used to add diversity of the dataset, meaning that different algorithms and comparisons that address more modes of micro-expressions can possibly be developed. Both the VIS and NIR subsets were collected with 25 fps and 640×480 resolution. In contrast to a down-sampled version of the data with 100 fps, VIS can be used to study the normal behavior of micro-expressions, e.g. when motion blurs in web camera footage appear. NIR images were photographed using a near-infrared camera and can be used to eliminate the effect of lighting illumination on micro-expressions. The drawbacks of the SMIC dataset pertain to its labelling methods: (1) only three emotion labels (positive, negative and surprise) are provided, (2) the emotion labeling was only based on participants' self-reporting, and given that different participants may rank the same emotion differently, this overall self-reporting may not be precise, and (3) the labels of action units (AUs) are not provided in SMIC.

AUs are fundamental actions of individual muscles or groups of muscles, which are defined according to the Facial Action Coding System (FACS) [38], a system used to categorize human facial movements by their appearance on the face. AUs are widely used to characterize the physical expression of emotions. Although FACS has successfully

TABLE 2: Seven publicly released micro-expression datasets.

Here, *samples** refers to the samples of micro-expressions. Note that CAS(ME)^2 and MMEW also contain samples of macro-expressions. Not all participants produce micro-expressions; here, *Participants** refers to the number of participants who generate micro-expressions. The parentheses enclose the number of emotional video clips in that category.

Characteristics	Datasets								
	MEVIEW	SMIC			CASME	CASME II	CAS(ME)^2	SAMM	MMEW
		HS	VIS	NIR					
Num of samples*	40	164	71	71	195	247	57	159	300
Participants*	16	16	8	8	35	35	22	32	36
Frame rate	25	100	25	25	60	200	30	200	90
Mean age	N/A	N/A			22.03	22.03	22.59	33.24	22.35
Ethnicities	N/A	3			1	1	1	13	1
Resolution	1280×720	640×480			640×480 & 1280×720	640×480	640×480	2040×1088	1920×1080
Facial resolution	N/A	190×230			150×190	280×340	N/A	400×400	400×400
Emotion classes	7 categories: Happiness (6) Anger (2) Disgust (1) Surprise (9) Contempt (6) Fear (3) Unclear (13)	3 categories: Positive (107) Negative (116) Surprise (83)			8 categories: Amusement (5) Disgust (88) Sadness (6) Contempt (3) Fear (2) Tense (28) Surprise (20) Repression (40)	5 categories: Happiness (33) Repression (27) Surprise (25) Disgust (60) Others (102)	4 categories: Positive (8) Negative (21) Surprise (9) Others (19)	7 categories: Happiness (24) Surprise (13) Anger (20) Disgust (8) Sadness (3) Fear (7) Others (84)	7 categories: Happiness (36) Anger (8) Surprise (89) Disgust (72) Fear (16) Sadness (13) Others (102)
Available labels	Emotion/FACS	Emotion			Emotion/FACS	Emotion/FACS	Emotion/FACS/ Video type	Emotion/FACS	Emotion/FACS
Download URL	http://cmp.felk.cvut.cz/cechj/ME/	http://www.cse.oulu.fi/SMIC Database			http://fu.psych.ac.cn/CASME/casme-en.php	http://fu.psych.ac.cn/CASME/casme2-en.php	http://fu.psych.ac.cn/CASME/cas(me)2-en.php	http://www2.docm.mmu.ac.uk/STAFF/M.Yap/dataset.php	http://www.dpailab.com/database.html

Fig. 3: Snapshots of six micro-expression datasets. From left to right and from top to bottom, these are as follows: MEVIEW [28], SAMM [29], SMIC [30], CASME [31], CASME II [32], and CAS(ME)^2 [33].

TABLE 3: Emotional facial action coding system for macro-expressions

Emotion	Action units
Happiness	6+12
Sadness	1+4+15
Surprise	1+2+5B+26
Fear	1+2+4+5+7+20+26
Anger	4+5+7+23
Disgust	9+15+16
Contempt	R12A+R14A

been used to code emotion-related facial action in macro-expressions (Table 3), the AU labels in different emotion categories of micro-expressions are diverse and consequently deserving of study in their own right. Different from SMIC, AU labels were provided in CASME, CASME II, CAS(ME)^2 and SAMM.

CASME, CASME II and CAS(ME)^2 were developed by the same group and utilized the same experimental protocol. In a well-controlled laboratory environment, four lamps were chosen to provide steady and high-intensity illumination (this approach can also effectively avoid the flickering of lights caused by alternative current). To elicit micro-expressions, participants were instructed to maintain a neutral facial expression when watching video episodes with high emotional valence. CASME II is an extended version of CASME, with the major differences between them

being as follows:

- CASME II used a high-speed camera with a sampling rate of 200 fps, while the sampling rate in CASME is only 60 fps;
- CASME II has a larger face size (of 280×340 pixels) in image sequences, while the face size in the CASME samples is only 150×190 pixels;
- CASME and CASME II have 195 and 247 micro-expression samples respectively. CASME II has a more uniformly balanced number of samples across each emotion class, i.e., Happiness (33 samples), Repression (27), Surprise (25), Disgust (60) and Others (102); by contrast, the CASME samples are much more poorly distributed, with some classes having very few samples (e.g., only 2 samples in fear and 3 samples in contempt; see Table 2).

Successfully eliciting micro-expressions by maintaining neutral faces is a difficult task in itself; it is even more difficult to elicit both macro- and micro-expressions. CAS(ME)^2 collected this kind of data, in which macro- and micro-expressions are distinguished by duration time; i.e., whether or not it is smaller than 0.5 seconds. Figure 4 presents macro- and micro-expressions of the same participant from the CAS(ME)^2 dataset. The CAS(ME)^2 dataset was divided into Part A and Part B: Part A contains 87 long videos with both macro and micro-expressions, while Part B contains 300
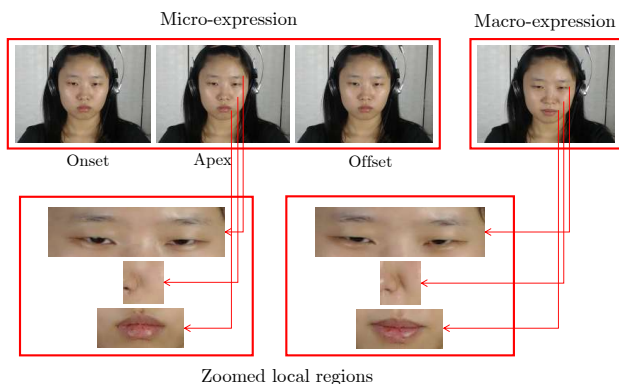
Fig. 4: Micro and macro-expression for the same person on the CAS(ME)^2 dataset. The first row is the original images, while the second row corresponds to the zoomed local regions of the micro-expression apex and the macro-expression of the same expression.

TABLE 4: The micro-expression categories and AU labeling

Emotion categories	Main AU labeling type	Quantity
Happiness	AU6/AU12/AU6+AU12	36
Surprise	AU5/AU26/AU1+AU2/AU18	89
Anger	AU17+AU13/AU7/AU9	8
Disgust	AU9/AU10/AU4+7/AU4+9	72
Fear	AU4+5/AU20/AU4+11/AU14	16
Sadness	AU14/AU17/AU14+17/AU1+2	13
Others	AU4/AU18/AU4+14/AU38/AU4+7	66

macro-expression samples and 57 micro-expression samples. The emotion classes includes positive, negative, surprise and others. All samples in CASME, CASME II and CAS(ME)^2 were coded into onset, apex and offset frames, and labeled based on a combination of AUs, the emotion types of emotion-evoking videos and self-reported emotion.

The SAMM dataset [29] contains 159 samples (i.e., image sequences containing spontaneous micro-expressions) recorded by a high-speed camera with 200 fps and a resolution of 2040×1088. Similar to the series of CASMEs, these samples were recorded in a well controlled laboratory environment with carefully designed lighting conditions, such that light fluctuations (which resulted in flickering on the recorded images) can be avoided. A total of 32 participants (16 males, 16 females, mean age 33.24) with a diversity of ethnicities (including 13 races) were recruited for the experiment. To assign emotion labels, in addition to self-reporting, each participant was required to fill in a questionnaire before starting the experiment, after which each emotional stimuli video was specially tailored for different participants to elicit the desired emotions in micro-expressions. Seven emotion categories were labelled in SAMM: contempt, disgust, fear, anger, sadness, happiness and surprise.

So far, CAS(ME)^2 is the most appropriate for micro-expression spotting, since it is close to the real scenarios in that it contains both macro- and micro-expressions. However, it is not suitable for micro-expression recognition, since the number of micro-expression samples is small (only 57). Among SMIC, CASME, CASME II and SAMM, SAMM is the best for micro-expression recognition, but also has the limitation of small sample numbers. Accordingly, to both

provide a suitable dataset for micro-expression recognition and inspire new research directions, we introduce a new dataset *micro-and-macro expression warehouse* (MMEW).

MMEW follows the same elicitation paradigm used in SMIC, CASME, CASME II, SAMM and CAS(ME)^2, i.e., watching emotional video episodes while attempting to maintain a neutral expression. The full details of the MMEW dataset construction process are presented in the supplemental material. All samples in MMEW were carefully calibrated by experts with onset, apex and offset frames, and the AU annotation from FACS was used to describe the facial muscle movement area (Table 4). Compared with the state-of-the-art CASME II, the major advantages of MMEW are as follows:

- The samples in MMEW have a larger image resolution (1920×1080 pixels), while the resolution of samples in CASME II is 640×480 pixels. Furthermore, MMEW has a larger face size in image sequences of 400×400 pixels, while the face size in the CASME II samples is only 280×340 pixels.
- MMEW and CASME II contain 300 and 247 micro-expression samples, respectively. MMEW has more elaborate emotion classes, i.e. Happiness (36), Anger (8), Surprise (89), Disgust (72), Fear (16), Sadness (13) and Others (66); by contrast, the emotion classes of Anger, Fear and Sadness are not included in CASME II. Furthermore, the class of Others in CASME II contains 102 samples (41.3% of all samples).
- 900 macro-expression samples with the same class category (Happiness, Anger, Surprise, Disgust, Fear, Sadness), acted out by the same group of participants, are also provided in MMEW (see Figure 1 for an example). These may be helpful for further cross-modal research (e.g., from macro- to micro-expressions).

Compared with the state-of-the-art SAMM, MMEW contains more samples (300 vs. 159; see Table 2). Moreover, MMEW contains sufficient samples of both macro- and micro-expressions from the same subjects (Figure 1). This new feature also opens up a promising new avenue of pre-training with macro-expression data from the own dataset instead of looking for other datasets, which we summarize in Section 8.3. MMEW is publicly available[1].

In Section 8, we use CAS(ME)^2 to evaluate micro-expression spotting methods, and further use SAMM and MMEW to evaluate micro-expression recognition, which can serve as baselines for the comparison of new methods developed in the future.

4 MICRO-EXPRESSION FEATURES

Since a micro-expression is a subtle and indiscernible motion of human facial muscles, the effectiveness of micro-expression spotting and recognition relies heavily on discriminative features that can be extracted from image sequences. In this section, we review existing features using a hierarchical taxonomy. Note that state-of-the-art deep learning techniques have been applied to micro-expression analysis in an end-to-end manner, so that hand-crafted features

1. http://www.dpailab.com/database.html

are less necessary. The related deep learning techniques will be reviewed in Section 6.2.

Throughout this section, all summarized features are computed on a preprocessing and normalized image sequence. The preprocessing involves detecting the facial region, identifying facial feature points, aligning the facial regions in each frame to remove head movements, normalizing or interpolating additional images into the original sequence, etc. The reader is referred to Sections 3.1 and 3.2 in [12] for more details on preprocessing.

The classical voxel representation depicts a given image sequence in a spatiotemporal space $\mathbb{R}^3$, with three dimensions x, y and t: here, x and y are pixel locations in one frame and t is the frame time. At each voxel $v(x, y, t)$, gray or color values are assigned. Subsequently, the local, tiny transient changes of facial regions within an image sequence can be effectively captured by local patterns of gray or color information at each voxel, which we refer to as a kind of *dynamic texture* (DT). DT has been widely studied in the computer vision field and is traditionally defined as image sequences of moving objects/scenes that exhibit certain stationarity properties in time (see e.g. [39]). Here, we treat DT as an extension of texture — local patterns of color at each image pixel — from the spatial domain to the spatiotemporal domain in an image sequence. There are three ways to make use of this local intensity pattern information:

- DT features in the original spatiotemporal domain $(x, y, t) \in \mathbb{R}^3$ (Section 4.1);
- DT features in the frequency domain: by applying Fourier or wavelet transforms to a signal in $\mathbb{R}^3$, the information in the spatiotemporal domain can also be dealt with in the frequency domain (Section 4.2);
- DT features in a transformed domain by tensor decomposition: by representing a color micro-expression image sequence as a 4th-order tensor, all the information in $(x, y, t) \in \mathbb{R}^3$ can be interpreted via tensor decomposition: namely, the facial spatial information is mode-1 and mode-2 of the tensor, temporal information is mode-3 and color information is mode-4 (Section 4.3);
- Optical flow features, which indicate the patterns of motion of objects (facial regions in our case) and are often computed by the changing intensity of pixels between two successive image frames over time, based on partial derivatives of the image signals (Section 4.4).

We examine these four feature classes in more detail in the following four subsections and all presented features are summarized in Table S2 in the supplementary material.

4.1 DT features in the spatiotemporal domain

4.1.1 LBP-based features

Many DT features in the spatiotemporal domain are related to *local binary patterns from three orthogonal planes* (LBP-TOP) [40]. The local binary pattern (LBP) [41] on each plane is represented as

$$LBP_{P,r} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p, \tag{1}$$

where c is a pixel at frame t, referred to the center, g_c is the gray value of c, g_p denotes the gray value of the pth pixel, P is the number of neighboring points located inside the circle of radius r centered at c, and $s(x)$ is an indicator function

$$s(x) = \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases} \tag{2}$$

By concatenating the LBPs' co-occurrence statistics in three orthogonal planes, Zhao et al. [40] proposed the novel LBP-TOP descriptor, which concatenates LBP histograms from three planes. LBP-TOP has been successfully applied to the recognition of both macro- and micro-expressions [30], [40].

Rotation-invariant features, such as LBP-TOP, do not fully consider all directional information. Ben et al. [42] proposed hot wheel patterns (HWP), HWP-TOP, and dual-cross patterns from three orthogonal planes (DCP-TOP) based on DCP [43]. The experiments in [42] show that these features, when enhanced by directional information description, can further improve the micro-expression recognition accuracy.

To resolve the issue that LBP-TOP may be coded repeatedly, Wang et al. [44] proposed an LBP with six intersection points (LBP-SIP), which can delete the six intersections of repetitive coding. In doing so, it reduces redundancy and the histogram length, and also improves the speed. Moreover, in order to preserve the essential patterns so as to improve recognition accuracy and reduce redundancy, Wang et al. [45] proposed the super-compact LBP-three mean orthogonal planes (MOP) for micro-expression recognition. However, the accuracy of this approach is slightly worse when dealing with short video. To make the features more compact and robust to changes in light intensity, Huang et al. [46] proposed a completed local quantization patterns (CLQP) approach that decomposes the local dominant pattern of the central pixel with the surrounding pixels into sign, magnitude and orientation respectively, then transforms it into a binary code. They also extended CLQP into 3D space, referred to as spatiotemporal CLQP (STCLQP) [47].

4.1.2 Second-order features

This class of features make use of second-order statistics to characterize new DT representations of micro-expressions [48]. Hong et al. [49] proposed a second-order standardized moment average pooling (called 2Standmap) that calculates low-level features (such as RGB) and intermediate local descriptors (such as Histogram of Gradient Orientation; HIGO) for each pixel by means of second-order average and max operations.

John et al. [50] proposed the re-parametrization of the second-order Gaussian jet for encoding LBP. This approach obtains a set of three-dimensional blocks; based on these, more robust and reliable histograms can be generated, which are suitable for different facial analysis tasks. Kamarol et al. [51] proposed a spatiotemporal texture map (called STTM) to convolve an input video sequence with a 3D Gaussian kernel function in a linear space representation. By calculating the second-order moment matrix, the spatiotemporal texture and the histogram of the micro-expression sequence are obtained. This algorithm captures subtle spatial and temporal variance in facial expressions

with lower computational complexity, and is also robust to illumination variations.

4.1.3 Integral projection

Integral projection — which is a one-dimensional curved shape pattern — have also been investigated in the micro-expression analysis context. Integral projection can be represented by a vector in which each entry corresponds to a 1D position (obtained by projecting the image along a given direction) and the entry value is the sum of the gray values of the projected image pixels at this position. Huang et al. [52] proposed a spatiotemporal LBP with Integral Projection (STLBP-IP); according to this approach, LBP is evaluated on the integral projections of image pixels obtained in the horizontal and vertical directions.

The same research group [52] also developed a revisited integral projection algorithm to maintain the shape property of micro-expressions, followed by LBP operators to further describe the appearance and motion changes from horizontal and vertical integral projections. They then proposed a discriminative spatiotemporal LBP with revisited integral projection (DiSTLBP-RIP) for micro-expression recognition [53]. Finally, they used a new feature selection method that extracted discriminative information based on the Laplacian method.

4.1.4 Other miscellaneous features

In addition to the aforementioned features, Polikovsky et al. [54] proposed a 3D gradient descriptor that employs AAM [55] to divide a face into 12 regions. By calculating and quantizing the gradient in all directions of each pixel, then constructing a 3D gradient histogram in each region, this descriptor extends the plane gradient histogram to capture the correlation between frames. Another widely used feature is the histogram of oriented gradients (HOG) [56] incorporating the convolution operation and weighted voting. These authors also proposed a histogram of image gradient orientation (HIGO) [56], which uses a simple vote rather than a weighted vote. HIGO can maintain good invariance of the geometric and optical deformation of the image. Moreover, the gradient or edge direction density distribution can better describe the local image appearance and shape, while small head motions can be ignored without affecting the recognition accuracy. HIGO is particularly suitable for occasions where the lighting conditions vary widely.

To maintain the invariance to both geometric and optical deformations of images, Chen et al. [57] employed weighted features and weighted fuzzy classification to enhance the valid information contained in micro-expression sequences; however, their recognition process is still expensive in terms of time cost. Lu et al. [58] proposed a Delaunay-based temporal coding model. This model divides the facial region into smaller triangular regions based on the detected feature points, then calculates the accumulated value of the difference between the pixel values of each triangular region in the adjacent frame by means of local temporal variations (LTVs). This approach encodes texture variations corresponding to facial muscle activities, meaning that any influence of personal appearance that is irrelevant to micro-expressions can be suppressed.

Wang et al. [59] used robust PCA (RPCA) to decompose a micro-expression sequence into dynamic micro-expressions with subtle motion information. More specifically, they utilized an improved local spatiotemporal directional feature (LSTD) [60] in order to obtain a set of directional codes for all six directions: i.e., XY(YX), XT(TX) and YT(TX). Subsequently, the decorrelated LSTD (DLSTD) was obtained by singular value decomposition (SVD) in order to remove the irrelevant information and thus allow the important micro-motion information to be emphasized. Zong et al. [61] designed a hierarchical spatial division scheme for spatiotemporal descriptor extraction, and further proposed a kernelized group sparse learning (KGSL) model to process hierarchical scheme-based spatiotemporal descriptors. This approach can effectively choose a good division grid for different micro-expression samples, and is thus more effective for micro-expression recognition tasks. Zheng et al. [62] developed a novel multi-task mid-level feature learning algorithm that boosts the discrimination ability of low-level features extracted by learning a set of class-specific feature mappings.

4.2 Frequency domain features

The micro-expression sequence can be transformed into the frequency domain by means of either Fourier or wavelet transforms. This makes relevant frequency features, such as amplitude and phase information, available for subsequent tasks. For example, the local geometric features (such as the corners of facial contours and the facial lines, which are easily overlooked by some feature description algorithms) can be easily identified by high-frequency information.

Oh et al. [63] extracted the magnitude, phase and orientation of the transformed image by means of Riesz wavelet transform. To discover the intrinsic two-dimensional local structures (i2D) of micro-expressions, Oh et al. [64] also performed a Fourier transform to restore the phase and orientation of the i2D via a high-order Riesz transformation using a Laplacian of Poisson (LOP) band-pass filter; this is followed by extracting LBP-TOP features and the corresponding feature histogram after quantification. An advantage of this approach is that the i2D can extract some easy-to-lose local structures, such as complex facial contours. Similar to i2D, the i1D [65] is extracted through the use of a first-order Riesz transformation. To obtain directional statistical structures from micro-expressions, Zhang et al. [66] used the Gabor filter to obtain the texture images of important frequencies and suppress other texture images.

Different from the aforementioned research, Li et al. [67] utilized Eulerian video magnification (EVM) to magnify the subtle motion in a video. In more detail, the representation of the frequency domain of the micro-expression sequence was obtained via Laplace transform, and some frequency bands were band-fed to enhance the corresponding scale of the movement and achieve directional amplification of the micro-expressions. Refining the EVM approach, Oh et al. [68] proposed the Eulerian motion magnification (EMM) method, which consists of amplitude-based EMM (A-EMM) and phase-based EMM (P-EMM). LBP-TOP was then used to extract the features of the micro-expression sequence, such that their algorithm enables the micro-movements to be magnified.

4.3 DT features in tensor-decomposition spaces

Treating the micro-expression sequence as a tensor enables rich structural spatiotemporal information to be extracted from it. Due to their high dimensionality, many tensor-based dimension reduction algorithms — which keep the inter-class distances as large as possible and the intra-class distances as small as possible — can be used to obtain a more effective discriminant subspace.

By viewing a gray-valued micro-expression sequence as a three-dimensional spatiotemporal tensor, Wang et al. [69] proposed a discriminant tensor subspace analysis (DTSA) that preserves some useful spatial structure information. More specifically, this method projects the micro-expression tensor to a low-dimensional tensor space in which the inter-class distance is maximized and intra-class distance is minimized.

Ben et al. [70] proposed a maximum margin projection with tensor representation (MMPTR) approach. This method also views a micro-expression sequence as a third-order tensor, and can directly extract discriminative and geometry-preserving features by maximizing the inter-class Laplacian scatter and minimizing the intra-class Laplacian scatter.

To obtain better discriminant performance, Wang et al. [71] proposed tensor independent color space (TICS). Representing a micro-expression sequence in three RGB channels, this approach extracted features from the four-order tensor space by utilizing LBP-TOP to estimate four projection matrices, each representing one side of the tensor data.

4.4 Optical flow features

Optical flow is a motion pattern of moving objects/scenes in an image sequence, which can be detected by the intensity change of pixels between two image frames over time. Many elegant optical flow algorithms have been proposed that are suitable for diverse application scenarios [72]. In micro-expression research, optical flow has been investigated as an important feature.

Patel et al. proposed the spatiotemporal integration of optical flow vectors (STIOF) [75], which computes optical flow vectors inside small local spatial regions and then integrates these vectors into the local spatiotemporal volumes. Liong et al. proposed an optical strain weighted features (OSWF) algorithm [83], which extracts the optical strain magnitude for each pixel and uses the feature extractor to form the final feature histogram. Xu et al. [84] proposed facial dynamics map (FDM) to capture small facial motions based on optical flow estimation [72].

Wang et al. [77] proposed a main directional maximal difference (MDMD) algorithm to characterize the magnitude of maximal difference in the main direction of optical flow features. More recently, Liu et al. proposed the Mean Directional Mean Optical Flow (MDMO) feature, which integrates the magnitude and direction of the main optical flow vectors from a total of 36 non-overlapping regions of interest (ROIs) in a human face [85]. While MDMO is a simple and effective feature, the average MDMO operation often loses the underlying manifold structure inherent in the feature space. To address this, the same research group further proposed a sparse MDMO [86] by constructing a dictionary containing all the atomic optical flow features in the entire video, as well as applying a time pool to achieve the sparse representation. MDMO and sparse MDMO are similar to the histogram of oriented optical flow (HOOF) feature [87]; however, the key difference is that MDMO and sparse MDMO use HOOF features in a local way (i.e., in each ROI region) and are thus more discriminative.

4.5 Deep features

In addition to the aforementioned hand-crafted features, deep learning methods that can automatically extract optimal deep features have also been applied recently to micro-expression analysis; we summarize this class of methods in Section 6.2. In these deep network models, the feature maps preceding the full connected layers (particularly the last full connected layer) can usually be regarded as deep features. However, these deep models are often designed as black boxes, and the interpretability or explainability of these deep features is frequently poor [88].

4.6 Feature interpretability

Compared with deep features which are implicitly learned through deep network optimization based on large training data, conventional hand-crafted features are usually designed based on either human experiences or statistical properties, which have better explainability. For examples, LBP-based features explicitly reflect a statistical distribution of local binary patterns. Integral projection method based on difference images can preserve the shape attributes of facial images. Optical-flow-feature-based methods use a kind of normalized statistic features that consider both local statistic motion information and spatial locations.

5 SPOTTING ALGORITHMS

In the literature, micro-expression detection and spotting are two related but often confused terminologies. In our study, we define micro-expression detection as the process of identifying whether or not a given video clip (or an image sequence) contains a micro-expression. Moreover, we put emphasis on micro-expression spotting, which goes beyond detection: in addition to detecting the existence of micro-expressions, spotting also identifies three time spots, i.e. onset, apex and offset frames, in the whole image sequence:

- the onset is the first frame at which a micro-expression starts (i.e., changing from the baseline, which is usually the neutral facial expression);
- the apex is the frame at which the highest intensity of the facial expression is reached;
- the offset is the last frame at which a micro-expression ends (i.e., returning back to the neutral facial expression).

In the early stages of micro-expression research, manual spotting based on the FACS coding system was used [38]. However, manual coding is laborious and time-consuming; e.g., recording a one-minute micro-expression video sample takes two hours on average [89]. Moreover, manual coding is subjective due to differences in the cognitive ability and living background of the participants [90]; it is therefore

TABLE 5: Comparison of representative micro-expression spotting algorithms

	References	Detection/spotting results	Advantages	Disadvantages
Optical flow based methods	[36] [73]	Distinguishing macro-expression from micro-expression	Able to capture tiny expression variations	A threshold is manually set, when the training data is small.
	[74]	Micro-expression detection/spotting		
	[75] [76]	Onset, apex and offset frames detection		
	[77]	Spotting facial movements from long-term videos	Able to obtain more accurate features of movement	Time-consuming
Feature-descriptor-based methods	[23] [78]	Onset, apex and offset frames detection	Simple	Only suitable for posed micro-expression, not spontaneous micro-expression
	[79] [56]	Micro-expression spotting	Satisfactory results	Very complicated; parameters and thresholds are manually set
	[80]	Micro-expression spotting	Error due to head motion is minimized	A threshold is not easy to determine
	[67]	Onset, apex and offset frames detection	Combination of two different features is more powerful	
	[81] [82]	Apex frame detection	Parameters and thresholds are set automatically	Only able to spot the apex frame

highly desirable to develop automatic algorithms for spotting micro-expressions.

There are three major challenges associated with developing accurate and efficient micro-expression spotting algorithms. First, detecting micro-expressions usually relies on setting the optimal upper and lower thresholds for any given feature (see Section 4:) the upper threshold aims at distinguishing micro-expressions from macro-expressions, while the lower threshold defines the minimal motion amplitude of micro-expressions. Second, different people may perform different extra facial actions. For example, some people blink habitually, while other people sniff more frequently, which can cause movement in the facial area. The impact of these facial motion areas on expression spotting should thus be taken into consideration. Third, when recording videos, many comprehensive factors (including head movement, physical activity, recording environment, and lighting) may significantly influence the micro-expression spotting.

Existing automatic micro-expression spotting algorithms can be broadly divided into two classes: namely, optical-flow-based and feature-descriptor-based methods. For ease of reading, we highlight the pros and cons of these two classes in Table 5.

5.1 Optical-flow-based spotting

Optical flow algorithms can be used to measure the intensity change of image pixels over the time. Shreve et al. [36], [73] proposed a strain pattern that was used as a measure of motion intensity to detect micro-expressions. However, this method relies on manually selecting thresholds to distinguish between macro-expression and micro-expression. Furthermore, this method was designed to detect posed micro-expressions, but not spontaneous micro-expressions (note that posed and spontaneous micro-expressions vary widely in terms of facial movement intensity, muscle movement, and time intervals). Shreve et al. [74] used optical flow to exploit non-rigid facial motion by capturing the optical strains. This method can achieve an 80% true positive rate with a 0.3% false positive rate in spotting micro-expressions. Moreover, this method can also plot the strains and visualize a micro-expression as it occurs over the time.

Patel et al. [75] used a discriminative response map fitting (DRMF) model [102] to locate key points of the face based on the FACS system. This method then groups the motion vectors of key points (indicated by the optical flow) and computes a cumulative value of the motion amplitude, shifted over time, to detect the onset, apex and offset frames of a micro-expression. However, this method also requires a manually specified threshold. Wang et al. [77] used the magnitude maximal difference in the main direction of the optical flow features to spot micro-expressions. Guo et al. [76] proposed a magnitude and angle combined optical flow feature for micro-expression spotting, and obtained more accurate results than the method proposed in [77].

5.2 Feature descriptor based spotting

Polikovsky et al. [23], [78] proposed to use the gradient histogram descriptor and K-Means algorithm to locate the onset, apex and offset frames of posed micro-expressions. While their method is simple, it only works for posed micro-expressions (not for spontaneous micro-expressions). Moilanen et al. [79] divided the face into 36 regions, calculated the LBP histogram of each region, then used the Chi-square distance between the current frame and the average to determine the degree of change in the video. While this method is novel, the design concept is somewhat complicated and the parameters in this method also need to be set manually.

Davison et al. [56] aligned and cropped faces for each video. This method involves splitting these faces into blocks and calculating the HOG for each frame. Afterwards, this method used Chi-Squared distance to calculate the dissimilarity between frames at a set interval in order to spot micro-expressions. Xia et al. [80] modeled the geometric deformation using an active shape model with SIFT descriptor [103] to locate the key points. Their method then performed the Procrustes transform between each frame and the first frame to remove the bias caused by the head motion. Subsequently, this method evaluated an absolute feature and its relative feature and combined them to estimate the transition probability. Finally, the micro-expression was detected according to a threshold. This method can minimize the error caused by head movement, but an optimal threshold is difficult to determine.

Li et al. [67] made use of the Kanade-Lucas-Tomasi algorithm [104] to track three specific points of every frame. This method extracted LBP and HOOF features from each

TABLE 6: Micro-expression recognition algorithms

	Algorithm	Advantages	Disadvantages
Traditional classifiers	SVM [71] [84] [85]	Most commonly used	Restricted by the limited training samples
	ELM [91] [59] [69]	Obtains an optimal solution	Restricted by the limited training samples
	KNN [87] [92]	Simple to use	Classification performance needs to be further improved
Deep learning	Selective CNN [93]	Avoids irrelevant deep features	The fitness function is subjective to overfitting
	DTSCNN [94]	Avoids the overfitting problem	Highly dependent on hardware
	DWLD +DBN [95]	Reduces the unnecessary learning for the redundant features	Insufficient samples
	TLCNN [96]	Solves the problem of limited micro-expression samples	Complicated
	ELRCN [97]	Solves the problem of limited micro-expression samples	There are shortcomings in preprocessing techniques and data augmentation
Transfer learning from other domains	DR [98]	Handles the unsupervised cross-dataset micro-expression recognition problem	The consistent feature distribution between the training and testing samples is seriously broken
	Source domain targetized [99]	Utilizes efficient speech data to enhance micro-expression recognition accuracy	Large difference of data distribution
	SVD [100] Coupled metric learning [42]	Solves the limited samples problem	Based on an assumption that a linear transformation exists between macro- and micro-expressions
	Auxiliary set selection model + a transductive transfer regression model [101]	Selects a small number of representative samples from the target domain data, rather than choosing all of them	Difficult to select the best sample number

area and obtained the differential value of every frame, then detected the onset, apex and offset frames based on a threshold. This method successfully fuses two different features to get more discriminative information; however, again, the threshold is difficult to determine.

Yan et al. [81] and Han et al. [105] proposed to using feature difference to locate the apex frame of a micro-expression. This method [81] used a constrained local model (CLM) [106] to locate 66 key points in a face, and then divided the face into subareas relative to these key points. Subsequently, they calculated the correlation between the LBP histogram feature vectors of every frame and first frame. Finally, the frame with the highest correlation value was regarded as the apex frame of a micro-expression. Different from the above methods, Liong et al. [82] spotted the apex frame by utilizing a binary search strategy and restricting the region of interest (ROI) to a predefined facial sub-region; here, the ROI selection was based on the landmark coordinates of the face. Meanwhile, three distinct feature descriptors, namely, CLM, LBP and Optical Strain (OS), were adopted to further confirm the reliability of the proposed method. Although the methods in [81] and [82] did not require the manual setting of parameters and thresholds, they are only able to spot the apex frame. Some disadvantages of the above-mentioned feature-descriptor-based methods include their high computational cost and the instability caused by noise and illumination variation.

6 RECOGNITION ALGORITHMS

Micro-expression recognition usually comprises two steps: feature extraction and feature classification. Traditional classification methods rely on artificially designed features (e.g., those summarized in Section 4). State-of-the-art deep learning methods can automatically infer an optimal feature representation and offer an end-to-end classification solution. However, deep learning requires a large number of training samples, while all existing micro-expression datasets are small; therefore, transfer learning that makes use of the knowledge from a related domain (in which large

datasets are available) has also been considered in micro-expression recognition. In this section, we summarize these three classes of recognition methods: traditional methods, deep learning and transfer learning methods. For ease of reading, we highlight the pros and cons of these methods in Table 6.

6.1 Traditional classification algorithms

Pattern classification/recognition has a long history [107], and many well-developed classifiers have already been proposed; the reader is referred to [108] for more details. In the special application of micro-expression recognition, many classic classifiers have been applied, including the support vector machine (SVM) [71], [84], [85], extreme learning machine (ELM) [91], [59], [69] and K nearest neighbor (KNN) classifier [87], [92], to name only a few.

6.2 Deep learning

In recent years, deep learning approaches that integrate automatic feature extraction and classification in an end-to-end manner have been great success. These deep models can obtain state-of-the-art predictive performance in many applications including micro-expression recognition [109], [110], [111].

Hao et al. [95] proposed an efficient deep network for micro-expression recognition. This network involves a two-stage strategy: the first stage use a double Weber local descriptor for extracting initial local texture features, and the second stage uses a deep belief net (DWLD+DBN) for global feature extraction. Wang et al. [96] proposed a transferring long-term convolutional neural network (TLCNN), which used Deep CNN to extract features from each frame of micro-expression video clips. Khor et al. [97] proposed an Enriched Long-term Recurrent Convolutional Network (ELRCN) that operates by encoding each micro-expression frame into a feature vector via CNN. It then predicts the micro-expression by passing the feature vector through a Long Short-term Memory (LSTM) module. Through the enrichment of the spatial dimension (via input channel

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/TPAMI.2021.3067464, IEEE Transactions on Pattern Analysis and Machine Intelligence

SUBMITTED TO IEEE T-PAMI

12

stacking) and the temporal dimension (via deep feature stacking), TLCNN and LSTM can effectively deal with the problem of small sample size.

To tackle the overfitting problem, Patel et al. [93] developed a selective deep model that can remove irrelevant deep information from micro-expression recognition. Their model has a good generalization ability. Peng et al. [94] further proposed a dual temporal scale convolutional neural network (DTSCNN) for micro-expressions recognition. DTSCNN used different streams to adapt to different frame rates of micro-expression video clips, each of which included an independent shallow network to prevent overfitting. These shallow networks were fed with optical-flow sequences to ensure that higher-level features could be extracted.

Since the deep network models have a huge number of parameters and weights, sufficiently well-labelled micro-expression samples are urgently needed in these models to improve the recognition performance.

6.3 Transfer learning from other domains

Due to the small data size of existing micro-expression datasets, transfer learning — which transfers knowledge from a source domain (in which a large sample size is available) to a target domain — has been considered in micro-expression recognition. Usually, the source and target domains are different but related (e.g., macro- and micro-expressions) so that they share some common knowledge (e.g., macro- and micro-expressions have similar AUs and DT information when expressing emotions) and benefitting from transferring this knowledge the performance in the target domain can be improved.

Zong et al. [98] proposed an effective framework, called domain regeneration (DR), for cross-dataset micro-expression recognition. The training and testing samples are derived from different micro-expression datasets. The DR framework is able to learn a regenerator that regenerates samples with similar feature distributions in source and target micro-expression datasets.

Another research group developed a series of transfer learning works [42], [99], [100] on micro-expression recognition. In [100], they proposed a macro-to-micro transformation model using singular value decomposition (SVD). This model takes advantage of sufficiently labelled macro-expression samples to increase the number of training samples. In [42], the authors extracted several local binary operators (that jointly characterize macro- and micro-expressions) and transferred these operators into a common subspace shared by source and target domains for learning purposes. In [99], they proposed to transfer the knowledge of speech samples in the source domain to the micro-expression samples in the target domain, in such a way that the transferred samples would have a similar feature distribution in the target domain. All these three transfer learning methods have been shown to achieve better recognition accuracy than some previous works.

Moreover, rather than transferring learning from macro-expressions or speech samples, a recent work [101] proposed an auxiliary set selection model (ASSM) and a transductive transfer regression model (TTRM) to bridge the gap between the source and target micro-expression domains. This method outperforms many state-of-the-art approaches, as the feature distribution difference between the source and target domains is small.

7 APPLICATIONS

Micro-expression analysis has a wide range of potential applications in the fields of criminal justice, business negotiation and psychological consultation, etc. In the below, we present the details of lie detection, which serves as a common and typical application scenario for micro-expression analysis.

One automatic lie detection instrument that is widely used today is the polygraph. This instrument usually collects multi-track physiology information such as respiration, pulse, blood pressure, pupil, skin electricity, and brain waves. If the test subject tells a lie, this instrument will record a fluctuation in some of the above physiological signals. However, whether the polygraph will work properly depends on many factors, such as the external environment during the test, the intelligence level of the tested person, the physical and mental condition of the tested person, and even the skill level of the polygraph operator. Sometimes, due to fear experienced as a result of the unfamiliar polygraph testing environment, some innocent people might show a state of panic and anxiety during the test, leading to false positives in the test results. On the other hand, some well-trained professionals may be able to pass a polygraph even while lying. In such a situation, micro-expression analysis can provide another method of lie detection.

Ekman [8] pointed out that as an effective clue for use when identifying lies, micro-expressions have broad applications in lie detection. By detecting the occurrence of micro-expressions and understanding the emotion type behind them, it will be possible to accurately identify the true intentions of the participant and improve the rate of success in lie detection. For example, when the participant reveals a micro-expression of happiness, this may indicate hidden delight at having successfully passed the test [4]. However, when the participant shows the micro-expression of surprise, this may indicate that the participant has never considered or understood the relevant questions. Because the micro-expression is difficult to hide and is usually the same as a person's true state of mind when lying, psychologists believe that the micro-expression is an important clue for lie detection. Pérez-Rosas et al. [112] discovered that the five micro-expressions most related to falsehood are: frowning, eyebrow raising, lip corners turning up, lips protruded and head turning to the side.

8 COMPARISON

The research on micro-expression features, spotting and recognition algorithms, such as that summarized in this paper, are scattered throughout the literature. A great challenge is that their performance results are reported across different experimental settings, making it difficult to conduct a fair comparison based on a common framework. In this section, we present a study that compares a selected set of representative spotting and recognition algorithms. As
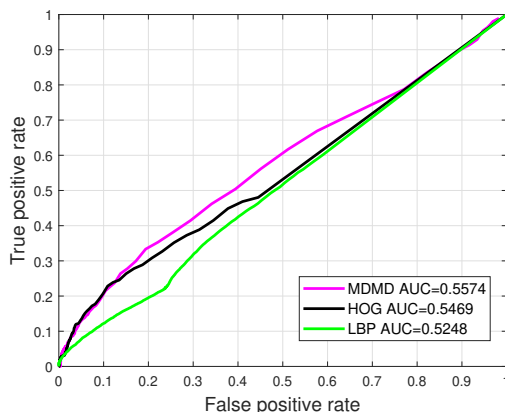
Fig. 5: ROC results of three methods — MDMD, HOG and LBP — on CAS(ME)^2 in terms of micro-expression spotting.
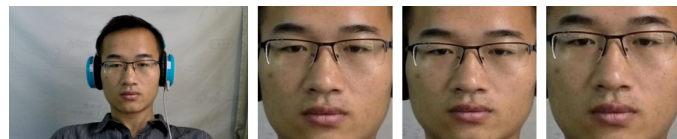

Fig. 6: Comparisons of three alignment algorithms. Due to the space limitations, we only show four frames here. From left to right: (a) an original image sequence; (b) results yielded by ASM+LWM [32]; (c) results yielded by DRMF+OFA [85]; (d) results yielded by JCFDA [113].

234 samples from 6 classes (i.e., happiness, surprise, anger, disgust, fear, sadness) were used[2]; in the SAMM dataset, 72 samples from 5 classes (i.e., happiness, surprise, anger, disgust, fear) were used[3]. In both datasets, all samples were randomly split into five subsets according to "subject independent" approach, and the number of subjects in each subset is equal; i.e., our random split scheme ensures there is no overlapped subject between the test and the training sets at the same time. Five-fold cross-validation was performed, after which the average recognition results were reported.

8.2.1 Evaluation of different preprocessing methods

Image sequence alignment and interpolation are two common preprocessing methods utilized in micro-expression recognition. We analyze the influences of these two methods by keeping the other modules unchanged. The details are summarized below.

Alignment algorithms. In the micro-expression recognition task, it is necessary for the micro-expression sequence to normalize the size of faces and align the face shapes across all of the different video samples. After all background parts are removed, only the facial areas in each video are preserved. More specifically, each image is normalized to a size of 231×231 pixels. Since the largest number of frames among all the image sequences is 108, the micro-expression image sequences are interpolated to the maximal value of 110 frames. Here, we evaluate three alignment algorithms, namely ASM+LWM [32], DRMF+Optical flow alignment (OFA) [85] and joint cascade face detection and alignment (JCFDA) [113], for face alignment on the MMEW dataset. The LBP-TOP is used as a baseline to evaluate these three alignment algorithms.

In the settings of LBP-TOP, the radii R_x, R_y of axes X and Y vary from 1 to 4. In order to avoid having too many parameter combinations, R_x is set to be equal to R_y, and the variation range of the radius R_t of the T axis is also ranged from 1 to 4. Thus, the numbers of neighborhood points of the XY, XT and YT planes are all set to be 8. The recognition rates of these three alignment algorithms are obtained based on the uniform pattern under different parameters and different radius configurations. Finally, an SVM classifier with RBF kernel [114] is applied. The experimental results are presented in Figures 6, 7 and S3 in supplemental material. From the figures, we can draw the following conclusions. First, compared with the other two

pointed out in Section 3, thus far, the CAS(ME)^2 dataset is most appropriate for spotting evaluation (Section 8.1), where the SAMM and MMEW datasets are most suitable for recognition evaluation (Section 8.2). By taking the advantage of the fact that it contains both macro- and micro-expressions of the same subjects, we also compared the experimental results by using different datasets for pre-training, and tested the performance on MMEW and SAMM (Section 8.3). By ensuring that all the factors (including data sample size and pre-processing) are the same, the analysis provided in this section can serve as a baseline for the evaluation of new algorithms in future work.

8.1 Evaluation of spotting algorithms

First, we study the influence of the feature extraction methods on the performance of micro-expression spotting. We use the MDMD [77], HOG [56] and LBP [79] methods for comparison in our spotting experiments on CAS(ME)^2. For MDMD and LBP, the micro-expression samples are divided into 5×5 and 6×6 blocks, respectively. For HOG, the number of blocks is 6×6, the signed gradient direction binning is set to 2π, and the number of direction bins is set to 8. Figure 5 plots the ROC curves of these three methods on CAS(ME)^2; the results show that MDMD performs the best in spotting micro-expressions. The possible reason is that MDMD applies the magnitude maximal difference in the main direction of optical flow features to detect/spot micro-expressions, leading to more notable differences or discriminant ability than those from the HOG or LBP features.

8.2 Evaluation of recognition algorithms

Compared to spotting, micro-expression recognition strongly depends on the amplification of subtle features, and then usually requires an additional preprocessing step. In Section 8.2.1, we first evaluate the effects of different preprocessing methods on the MMEW dataset. Then we perform a unified comparison on the MMEW and SAMM datasets to evaluate the recognition accuracies of traditional methods (using hand crafted features) in Section 8.2.2 and state-of-the-art methods (including deep learning methods) in Section 8.2.3, respectively.

Throughout this section, all recognition results were obtained under the following settings. In the MMEW dataset,

---

2. Because transfer learning methods were included in our comparison, we exclude 66 samples in the "Others" category.

3. We exclude 3 samples from the "Sadness" category (due to its small sample size) and 84 samples from the "Others" category due to the inclusion of transfer learning in our comparison.
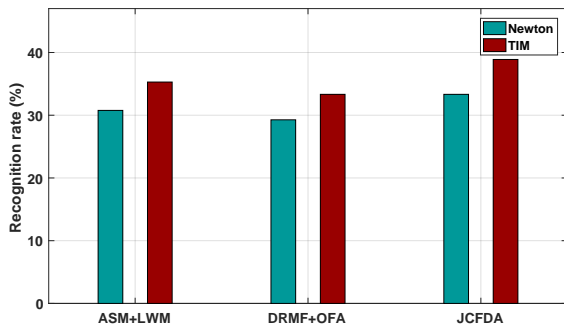
Fig. 7: The highest recognition rates of LBP-TOP with three alignment and two interpolation algorithms on MMEW.

TABLE 7: Recognition rates (%) of micro-expressions using hand-crafted features with different classifiers on MMEW and SAMM.

Hand-crafted features	MMEW			SAMM		
	KNN	SVM	ELM	KNN	SVM	ELM
LBP-TOP [40]	34.5	38.9	32.4	31.9	37.0	30.5
DCP-TOP [42]	37.6	42.5	36.2	36.1	36.8	32.8
LHWP-TOP [42]	39.4	43.2	36.2	40.5	41.7	36.1
RHWP-TOP [42]	37.6	45.9	37.8	36.1	38.1	43.8
LBP-SIP [44]	35.3	43.9	36.6	36.5	37.4	35.8
LBP-MOP [45]	43.9	41.5	34.6	32.4	35.3	34.7
STLBP-IP [52]	36.6	46.3	46.6	35.7	42.9	39.3
DiSTLBP-RIP [53]	39.0	44.0	41.5	42.9	46.2	42.9
FDM [84]	30.8	34.6	31.4	33.3	34.1	32.4
MDMO [85]	53.4	60.6	65.7	44.1	50.0	50.0
Sparse MDMO [86]	42.9	51.0	60.0	44.7	52.9	52.9

methods, JCFDA [113] achieves the best recognition rate of 38.9%. The reason is that JCFDA [113] combines face detection with alignment and learns at the same time under a cascade framework; this joint learning greatly improves the alignment and real-time performance. In the following experiments, we thus choose JCFDA [113] to preprocess the micro-expression image sequences.

Interpolation algorithms. Two interpolation algorithms, namely Newton interpolation [70] and TIM interpolation [115], are used to interpolate micro-expression sequences into different numbers of frames, including 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110 frames. The highest recognition rates of these two algorithms with different interpolated frame numbers are provided in Figure S4 in supplemental material. We can observe that as the number of frames increases, the recognition rates of the two interpolation algorithms increase initially and then decrease towards the end. Newton interpolation can obtain its highest recognition rate of 33.3% under the interpolated frame number of 60. Compared with Newton interpolation, TIM interpolation has relatively higher recognition rate, with the highest recognition rate reaching 38.9% when the micro-expression sequence is respectively interpolated to 30 or 60 frames. Accordingly, in the following experiments, TIM is used to interpolate each micro-expression sequence to 60 frames.

8.2.2 Comparisons of traditional methods

In this section, we evaluate the recognition performances of representative traditional methods that use hand-crafted features, specifically LBP-TOP [40], DCP-TOP [42], LHWP-TOP [42], RHWP-TOP [42], LBP-SIP [44], LBP-MOP [45], STLBP-IP [52], DiSTLBP-RIP [53], FDM [84], MDMO [85] and Sparse MDMO [86]. Meanwhile, we select KNN, SVM (with RBF kernel) and ELM [91] as three representative classifiers, due to the fact that these have been selected by most previous micro-expression recognition works. All of the above-mentioned methods follow their original settings outlined in the respective publications, except for STLBP-IP and DiSTLBP-RIP. Based on [52], we divide each frame into 5×5 blocks before feature extraction using STLBP-IP. In DiSTLBP-RIP, we produce the difference images of micro-expression image sequence, but we do not use robust principal component analysis (RPCA) [116] as in [53].

Table 7 lists the best recognition rates of these hand-crafted features combined with different classifiers on MMEW and SAMM. Except for LBP-TOP and sparse MDMO, the chosen parameters of each compared method

are the same on MMEW and SAMM. The sizes of radii R_x, R_y and R_t on the three orthogonal planes (YT, XT and XY) of DCP-TOP, LHWP-TOP, RHWP-TOP, LBP-SIP, LBP-MOP are $(2, 2, 4)$, $(4, 8, 4)$, $(5, 8, 4)$, $(1, 1, 2)$, and $(1, 1, 2)$, respectively. The parameter for LBP-TOP is $(1, 1, 4)$ on both MMEW and SAMM. For STLBP-IP, the frame number of each video clip is set to 45, and the size of the linear mask of the one-dimensional local binary pattern (1DLBP) is set to 9. Moreover, for DiSTLBP-RIP, each frame is also divided into 5×5 blocks, and the dimension of the discriminative group-based feature using feature selection is 95. For FDM, each frame is divided into 6×6 blocks. Sparse MDMO achieves its best performance with the following parameters on MMEW (and SAMM): the sparsity balance weight is 0.4 (and 0.4), the dictionary size is 256 (and 256), and the pooling parameter is 1 (and 0.1).

From Table 7, it can be observed that generally speaking, better recognition accuracies are obtained by either SVM or ELM, but the performance of ELM is relatively unstable. MDMO and Sparse MDMO have better recognition rates on both datasets. MDMO (65.7%, 50%) and sparse MDMO (60%, 52.9%) are the top two methods on both MMEW and SAMM. The reasons are as follows: MDMO and sparse MDMO are based on optical flow features extracted from ROI. Both methods utilize local statistical motion and spatial location information, and further apply the robust optical flow calculation method to capture the texture part of the image, followed by affine transformation. Therefore, the obtained optical flow field is insensitive to illumination conditions and head movement, which is important to have a good performance on SAMM.

8.2.3 Comparisons of state-of-the-art methods

We next evaluate the recognition performance of state-of-the-art methods (including deep learning methods, multi-task mid-level feature learning [62] and KGSL [61]) on MMEW and SAMM. Table 8 summarizes the comparison results. The handcrafted features in Table 7 are also kept in Table 8 for ease of comparison. We note that the ranking of these methods is almost consistent on both datasets.

For the multi-task mid-level feature learning [62], LBP-TOP, LBP-MOP and an extension of LBP-MOP are used as the low-level features. When compared with the performance of the original low-level features (such as LBP-

TABLE 8: Recognition rates (%) of micro-expressions using the state-of-the-art methods on MMEW and SAMM. The ranking of these methods are almost consistent on both datasets. (The best record of each dataset is marked in bold.)

Methods	Recognition rate (%)	
	MMEW	SAMM
FDM [84]	34.6	34.1
ResNet10 [117]	36.6	39.3
Handcrafted features + deep features [118]	36.6	47.1
LBP-TOP [40]	38.9	37.0
Selective deep features [93]	39.0	42.9
ELRCN [97]	41.5	46.2
DCP-TOP [42]	42.5	36.8
ESCSTF [119]	42.7	46.9
LHWP-TOP [42]	43.2	41.7
LBP-MOP [45]	43.9	35.3
LBP-SIP [44]	43.9	37.4
DiSTLBP-RIP [53]	44.0	46.2
RHWP-TOP [42]	45.9	38.1
STLBP-IP [52]	46.6	42.9
ApexME [120]	48.8	50.0
Transfer Learning [121]	52.4	55.9
Multi-task mid-level feature learning [62]	54.2	55.0
KGSL [61]	56.9	48.6
Sparse MDMO [86]	60.0	52.9
MDMO [85]	65.7	50.0
DTSCNN [94]	65.9	69.2
TLCNN [96]	**69.4**	**73.5**

TOP [40] and LBP-MOP [45]), multi-task mid-level feature learning [62] can enhance the discrimination ability of the original features [40], [45]. In KGSL [61], hierarchical STLBP-IP is used as the spatio-temporal descriptor, which benefits from a hierarchical spatial division scheme, so that KGSL [61] obtains the third best results among the non-deep-learning methods (inferior to MDMO and Sparse MDMO). Moreover, it is noticeable that all deep learning methods perform better than those utilizing handcrafted features. Setting up deep learning methods includes parameters, architectures, convolution layers, size of filters, batch normalization, etc.; the details are presented below.

Setting 1. The learning rate of ResNet10 [117] is 0.0001. The batch size is set to 8. ResNet10 includes 3 blocks and 2 fully connected layers; each block consists of two 3D convolution layers with convolution kernel of $3 \times 3 \times 3$ and one down-sampling layer with convolution kernel of $1 \times 1 \times 1$. The output sizes of the two fully connected layers are 128 and 6 respectively.

Setting 2. For handcrafted features + deep features [118], we employ two scales (10, 20 pixels) and three orientations (0, 60 and 120 degrees), resulting in 6 different Gabor filters, and set the sizes of radii on the three orthogonal planes of LBP-TOP to (1, 1, 2). The deep CNN contains 5 convolutional layers and 3 fully connected layers. The convolution kernel sizes are 11×11, 5×5, 3×3, 3×3 and 3×3, respectively. The output sizes of the 3 fully connected layers are 9216, 4096 and 1000, respectively.

Setting 3. ELRCN [97] is based on VGG-16, which contains 13 convolutional layers (3×3 conv) and 3 fully connected layers. The output sizes of 3 fully connected layers are 4096, 4096 and 2622, respectively. The output size of LSTM is 3000.

Setting 4. ApexME [120] is also based on VGG-16, with the same number of layers and convolution kernel size as ELRCN [97]. The output sizes of the 3 fully connected layers are 256, 64 and 6, respectively. The batch size is set to 128. During the fine-tuning, the drop-out rate is set to 0.8 in order to avoid over-fitting.

Setting 5. In TLCNN [96], the CNN contains 5 convolutional layers (11×11, 5×5, 3×3, 3×3, 3×3 conv) and 3 fully-connected layers. Moreover, there is a dropout layer after the first and second fully-connected layer. The learning rates for all training layers are all set to 0.001. The output sizes of the 3 fully connected layers are 256, 64 and 6, respectively.

Setting 6. In DTSCNN [94], the number of samples is extended to 90 for each class. All sequences are interpolated into different numbers of frames, i.e. 129 and 65. Their optical flow features are input into DTSCNN, which is a two-stream network containing 3D convolution and pooling units. The first network contains 4 convolutional layers ($3 \times 3 \times 8$, $3 \times 3 \times 3$, $3 \times 3 \times 3$, $3 \times 3 \times 4$ conv) and 1 fully connected layer, the output size of which is 6. Different from the first network, the convolution kernel sizes are $3 \times 3 \times 16$, $3 \times 3 \times 3$, $3 \times 3 \times 3$, and $3 \times 3 \times 4$.

Setting 7. For selective deep features [93], the CNN contains 10 convolutional layers ($3 \times 3 \times 3$ conv) and 2 fully-connected layers. After every two convolutional layers, there is a down-sampling layer ($1 \times 1 \times 1$ conv). In order to remove the irrelevant deep features trained by ImageNet and the facial expression dataset, evolutionary search is used to generalize the micro-expression data. The values for mutation probability, iteration number and population size are set to 0.02, 25 and 60, respectively.

Setting 8. The same ResNet10, LSTM and fully connected layers are used in transfer learning [121] and ESCSTF [119]. ResNet10 contains 3 blocks and 2 fully connected layers; moreover, each block consists of two 3D convolution layers ($3 \times 3 \times 3$ conv) and one down-sampling layer ($1 \times 1 \times 1$ conv). The output sizes are [8, 4, 256, 28, 28] respectively. The output of ResNet10 is fed to LSTM ($3 \times 3 \times 3$ conv). Finally, the output sizes of 2 fully connected layers are 1024 and 6. Although the two networks have the same structure, the training methods are different.

ResNet10 is pre-trained on ImageNet, then fine-tuned on macro- and micro-expression samples (16 interpolated frames for each sample) through transfer learning [121]. Moreover, in ESCSTF [119], 5 key frames are used to train ResNet10 for each micro-expression sample; namely, the onset, onset to apex transition, apex, apex to offset transition and offset frames.

ResNet10 [117], TLCNN [96], transfer learning [121] and ESCSTF [119] use batch-normalization, which is implemented before the activation function (for example, ReLU). The purpose is to normalize the data to a mean of 0 and a variance of 1.

In Table 8, it can be seen that TLCNN [96] achieves the best recognition performance (69.4% on MMEW and 73.5% on SAMM). This is due to (1) the use of the macro-expression samples[4] for pre-training, and (2) the use of

---

4. MMEW contains macro-expressions. For SAMM, the CK+ dataset [122] are used as macro-expression samples.
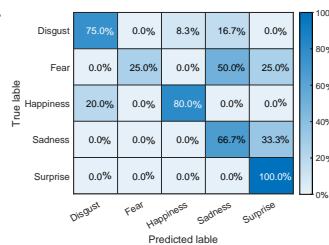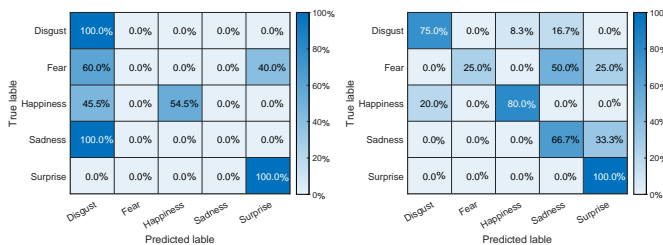
Fig. 8: Confusion matrices predicted by TLCNN on MMEW and SAMM. (a) MMEW; (b) SAMM.

the micro-expression samples fine-tuning, which solves the problem of the insufficient number of micro-expression samples. Moreover, LSTM is used to extract the discriminative dynamic characteristics from micro-expression sample sequences. We also present confusion matrices in Figure 8, from which we observe that in MMEW, all the "disgust" and "surprise" samples can be completely recognized on MMEW, instead the "fear" and "sadness" samples are difficult to train. Because about four fifths of fear (16) and sadness (13) samples of MMEW were used for fine-tuning (where the number in brackets is the total sample number of each class) and the samples for fine-tuning are too few. A similar situation occurs in SAMM where the total sample number of fear and sadness are 7 and 3, respectively. The classes for instance "fear" and "sadness" are more likely to be inconsistent in the classification results (Figure 8b).

Among deep learning methods, the baseline is ResNet10 [117], which can only obtain a recognition rate of 36.6% on MMEW and 39.3% on SAMM, due to over-fitting and the limited number of samples. The samples from the ImageNet dataset differ greatly from micro-expression samples, and a limited fine-tuning effect spoils the recognition performance; therefore, transfer learning [121] achieves recognition accuracies of 52.4% on MMEW and 55.9% on SAMM.

8.3 Encoding micro- from macro-expressions

The success of TLCNN [96] demonstrates that the macro-expressions are relevant for pre-training the network; this effectively alleviates the famous 3S (small sample size) problem inherent in micro-expression datasets. We further performed three group experiments: (1) CK+ was used for pre-training while MMEW (Micro) was applied for fine-tuning and testing, noted as "CK+ → MMEW (Micro)"; (2) MMEW (Macro) was used for pre-training while MMEW (Micro) was applied for fine-tuning and testing, noted as "MMEW (Macro) → MMEW (Micro)"; (3) CK+ was used for pre-training while SAMM was applied for fine-tuning and testing, noted as "CK+ → SAMM". The training and testing sets of MMEW were set as above-mentioned. Table 9 lists the data source and accuracy of the pre-training, fine-tuning and testing phases for each experiment. The results, i.e., MMEW (Macro) → MMEW (Micro) > CK+ → MMEW (Micro) (with descending effect), indicate that the knowledge of macro-expressions is useful for micro-expression recognition, and using macro- and micro-expressions from the same dataset (collected in the same laboratory environment) to pre-trained and fine-tune has better performance than using those from different datasets. This also verifies our motivation to establish MMEW, i.e. dataset containing both macro- and micro-expression of the same subject is

useful for transferring macro-expression knowledge to assist micro-expression recognition. The advantage of the new released MMEW dataset brings up an interesting problem:

"Do the macro-expressions of the same person help with recognizing his/her own micro-expressions?" Since MMEW contains the macro- and micro-expressions of the same participants, we also study this problem in a subject-dependent way. Note that although $CAS(ME)^2$ also contains the macro- and micro-expressions of the same participants, the number of micro-expression samples is small (only 57), so that $CAS(ME)^2$ is not suitable for studying this problem.

Accordingly, by utilizing MMEW, we conducted a preliminary experiment to evaluate different combinations of macro- and micro-expressions. (1) *Subject-independent evaluation:* The final average recognition rate was 69.4%. (2) *Subject-dependent evaluation:* All the micro-expression samples in MMEW were randomly divided into five folds, and for each fold, all macro-expressions in MMEW were used for pre-training. Then five-fold cross validation was also used for performance evaluation. The final average recognition rate was increased to 87.2%. These results demonstrate that macro-expressions of the same person are more relevant than those of different persons for pre-training the network. Therefore, MMEW can be used to explore new research directions, including subject-independent and subject-dependent encoding from macro-to micro-expressions.

9 FUTURE DIRECTIONS

Despite the significant progress made in micro-expression analysis over the last decade, several outstanding issues and new avenues exist for future development. Below, we propose some potential research directions.

Privacy-protection micro-expression analysis. Akin to macro-expressions, micro-expressions also constitute a kind of private facial information. Micro-expression spotting and recognition conducted by utilizing federated learning from decentralized data distributed across private user devices deserves study in its own right [123].

Utilizability of macro-expressions for micro-expression analysis. Both macro- and micro-expressions can be characterized according to the emotional facial action coding system. Considering the benefits offered by the new MMEW dataset, it would be interesting to explore the mutual effect of macro- and micro-expressions, particularly those from the same subject. Furthermore, as previous research has indicated that many deceptive behaviors are dependent on individual differences [124], we conjecture that micro-expression behavior is likely subject-dependent and MMEW dataset provides a new platform for both subject-dependent and subject-independent research in the future.

Standardized datasets. There are two major problems in the current micro-expression datasets. First, the existing number of micro-expression samples is too limited to facilitate proper training, since inducing micro-expressions is quite difficult. Researchers often require participants to watch emotional videos to elicit their emotions, and even to disguise their expressions. However, some participants may not exhibit micro-expressions under these circumstances, or may only exhibit them rarely. In addition, the

TABLE 9: Data sources and recognition rates (%) of the pre-training, fine-tuning and testing phases.

Experiments	Pre-training		Fine-tuning		Testing	
	Data source	Rec. rate	Data source	Rec. rate	Data source	Rec. rate
CK+ → MMEW (Micro)	CK+	99.9%	MMEW (Micro)	99.8%	MMEW (Micro)	65.6%
MMEW (Macro) → MMEW (Micro)	MMEW (Macro)	92.0%	MMEW (Micro)	96.6%	MMEW (Micro)	69.4%
CK+ → SAMM	CK+	99.0%	SAMM	99.7%	SAMM	73.5%

encoding/labeling of micro-expressions is time-consuming and laborious, since it requires the viewer to (1) watch the video at a slow speed and (2) select the onset, apex and offset frames of the facial motion, then calculate their duration. Consequently, there is no uniform standard available to annotate the emotion of micro-expressions. Second, owing to the poor quality of the videos in many existing micro-expression datasets, the full details of the low-intensity micro-expressions cannot be fully captured. Therefore, videos with higher temporal and spatial resolutions are needed for future algorithm design.

Data augmentation. Some data augmentation tricks used by the deep learning community could be helpful to increase the amount of available micro-expression data. For example, image rotation, translating, cropping and similar techniques will not change the labels of micro-expression but will increase the amount of available micro-expression data, leading to a potential performance improvement.

GAN-based sample generation. Deep learning methods achieve superior performance in facial expression recognition tasks. For example, Deng et al. [125] present an extensive and detailed review of state-of-the-art deep learning techniques for facial expression recognition. However, it is difficult to apply them to micro-expression recognition, since they will suffer due to the lack of sufficient training samples. In addition to manipulating images for data augmentation, it would be feasible to address this issue through utilizing generative adversarial networks (GANs) in order to generate a large number of pseudo-micro-expression samples, provided that we can define some criteria to ensure that the generated samples are indeed micro-expressions.

Multi-task learning. The extraction of micro-expression heavily depends on the ability to detect facial feature points, the semantic location of which has been predefined. The reason is that the motion amplitude of micro-expression is quite subtle, and facial feature points detection can reduce the effect caused by head movement during data preprocessing. It would therefore be useful in the future to design an end-to-end model capable of both learning the motion amplitude of micro-expression and detecting facial feature points. In this way, we could potentially alleviate the cost of micro-expression annotations.

Explainable micro-expression analysis. Although deep learning methods have received increasing attention and achieved good performance in micro-expression analysis, these deep models are usually treated as black boxes and have poor interpretability and explainability. In the critical applications such as lie detection and criminal justice, explainability is very important to help human understand the reasons behind predictions [126].

10 CONCLUSION

Micro-expression analysis has a wide range of potential real-world applications; for example, enabling people to

detect micro-expressions in daily life and develop a good interpretation/understanding of what lies behind micro-expressions. To make micro-expression analysis useful in practice, we need to develop robust algorithms with valid and reliable samples in order to make the spotting and recognition of micro-expressions applicable to real situations. Accordingly, in this survey, we review the current research on spontaneous facial micro-expression analysis (including datasets, features and algorithms) and propose a new dataset, MMEW, for micro-expression recognition. We further compare the performance of existing state-of-the-art methods, analyze the potential, and highlight the outstanding issues for future research on micro-expression analysis. Micro-expression analysis has recently become an active research area; accordingly, we hope this survey can help researchers, as a starting point, to review the developments in the state-of-the-art and identify possible directions for their future research.

REFERENCES

[1] K. R. Scherer, "What are emotions? and how can they be measured?" *Social Science Information*, vol. 44, no. 4, pp. 695–729, 2005.

[2] A. Freitas-Magalhães, "The psychology of emotions: The allure of human face," *University Fernando Pessoa Press, Oporto*, 2007.

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

[4] P. Ekman, "Emotions revealed: Recognizing faces and feelings to improve communication and emotional life." *Holt Paperback*, vol. 128, no. 8, pp. 140–140, 2003.

[5] P. Ekman and W. V. Friesen, "Nonverbal leakage and clues to deception," *Psychiatry-interpersonal & Biological Processes*, vol. 32, no. 1, pp. 88–106, 1969.

[6] B. Bhushan, "Study of facial micro-expressions in psychology," *Understanding Facial Expressions in Communication*, pp. 265–286, 2015.

[7] W. J. Yan, Q. Wu, J. Liang, Y. H. Chen, and X. Fu, "How fast are the leaked facial expressions: The duration of micro-expressions," *Journal of Nonverbal Behavior*, vol. 37, no. 4, pp. 217–230, 2013.

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

[9] S. Porter and L. T. Brinke, "Reading between the lies: Identifying concealed and falsified emotions in universal facial expressions," *Psychological Science*, vol. 19, no. 5, pp. 508–514, 2008.

[10] 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. Sheraton New York*, 2009.

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/TPAMI.2021.3067464, IEEE Transactions on Pattern Analysis and Machine Intelligence

SUBMITTED TO IEEE T-PAMI

18

[11] 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 Int. Conf. Acoust. Speec. Sign. Proc.*, 2019, pp. 2022–2026.

[12] 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.

[13] 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.

[14] W. E. Rinn, "The neuropsychology of facial expression: A review of the neurological and psychological mechanisms for producing facial expressions," *Psychological Bulletin*, vol. 95, no. 1, pp. 52–77, 1984.

[15] D. Matsumoto and H. S. Hwang, "Evidence for training the ability to read microexpressions of emotion," *Motivation and Emotion*, vol. 35, no. 2, pp. 181–191, 2011.

[16] P. Ekman and W. V. Friesen, "The repertoire of nonverbal behavior: Categories, origins, usage, and coding," *Semiotica*, vol. 1, no. 1, pp. 49–98, 1969.

[17] P. Ekman and E. Rosenberg, "What the face reveals : basic and applied studies of spontaneous expression using the facial action coding system (FACS)," *Oxford University Press*, vol. 68, no. 1, pp. 83–96, 2005.

[18] Q. Wu, X.-B. Sheng, and X. Fu, "Micro-expression and its applications," *Advances in Psychological Science*, vol. 18, no. 9, pp. 1359–1368, 2010.

[19] O. J. Rothwell J, Bandar Z, "Silent talker: A new computer-based system for the analysis of facial cues to deception," *Applied Cognitive Psychology*, vol. 20, no. 6, pp. 757–777, 2006.

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

[21] J. Wojciechowski, M. Stolarski, and G. Matthews, "Emotional intelligence and mismatching expressive and verbal messages: A contribution to detection of deception," *PLoS ONE*, vol. 9, no. 3, p. e92570, 2014.

[22] X. Zeng, Q. Wu, S. Zhang, Z. Liu, Q. Zhou, and M. Zhang, "A false trail to follow: differential effects of the facial feedback signals from the upper and lower face on the recognition of micro-expressions," *Frontiers in Psychology*, vol. 9, p. 2015, 2018.

[23] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," in *Crime Detection and Prevention (ICDP 2009), 3rd International Conference on*. IET, 2009.

[24] P. Ekman and M. O'Sullivan, "From flawed self-assessment to blatant whoppers: the utility of voluntary and involuntary behavior in detecting deception," *Behavioral Sciences & The Law*, vol. 24, no. 5, pp. 673–686, 2006.

[25] P. Ekman, "Darwin, deception, and facial expression," *Annals of the New York Academy of Sciences*, vol. 1000, no. 1, pp. 205–221, 2003.

[26] ——, "Lie catching and microexpressions," *The philosophy of deception*, pp. 118–133, 2009.

[27] C. M. Hurley and M. G. Frank, "Executing facial control during deception situations," *Journal of Nonverbal Behavior*, vol. 35, no. 2, pp. 119–131, 2011.

[28] J. M. Petr Husak, Jan Cech, "Spotting facial micro-expressions "in the wild"," in *In 22nd Computer Vision Winter Workshop*, 2017.

[29] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A spontaneous micro-facial movement dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, 2018.

[30] X. Li, T. Pfister, X. Huang, and G. Zhao, "A spontaneous micro-expression database: Inducement, collection and baseline," in *IEEE Int. Conf. Works. Automat. FG Recog.*, 2013, pp. 1–6.

[31] 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 Int. Conf. Works. Automat. FG Recog.*, 2013, pp. 1–7.

[32] 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.

[33] 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 Trans. Affect. Comput.*, vol. 9, no. 4, pp. 424–436, 2018.

[34] P. Ekman and W. V. Friesen, "Detecting deception from the body or face," *Journal of Personality and Social Psychology*, vol. 29, no. 3, pp. 288–298, 1974.

[35] M. Frank and P. Ekman, "The ability to detect deceit generalizes across different types of high-stake lies," *Journal of Personality and Social Psychology*, vol. 72, no. 6, pp. 1429–39, 1997.

[36] M. Shreve, S. Godavarthy, D. Goldgof, and S. Sarkar, "Macro-and micro-expression spotting in long videos using spatio-temporal strain," in *IEEE Int. Conf. Automat. FG Recog. Works.*, 2011, pp. 51–56.

[37] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. 59–69, 2009.

[38] P. Ekman and W. V. Friesen, "Facial action coding system (FACS): a technique for the measurement of facial actions," *Rivista Di Psichiatria*, vol. 47, no. 2, pp. 126–38, 1978.

[39] G. Doretto, A. Chiuso, Y. N. Wu, and S. Soatto, "Dynamic textures," *International Journal of Computer Vision*, vol. 51, no. 2, pp. 91–109, 2003.

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

[41] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 7, pp. 971–987, 2002.

[42] 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.

[43] C. Ding, J. Choi, D. Tao, and L. S. Davis, "Multi-directional multi-level dual-cross patterns for robust face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 38, no. 3, pp. 518–531, 2016.

[44] Y. Wang, J. See, R. C.-W. Phan, and Y.-H. Oh, "LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition," in *Asian Conference on Computer Vision*, 2014, pp. 525–537.

[45] ——, "Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition," *PLoS ONE*, vol. 10, no. 5, p. e0124674, 2015.

[46] X. Huang, G. Zhao, X. Hong, M. Pietikäinen, and W. Zheng, "Texture description with completed local quantized patterns," in *Scandinavian Conference on Image Analysis*, 2013, pp. 1–10.

[47] X. Huang, G. Zhao, X. Hong, W. Zheng, and M. Pietikäinen, "Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns," *Neurocomputing*, vol. 175, pp. 564–578, 2016.

[48] M. Niu, J. Tao, Y. Li, J. Huang, and Z. Lian, "Discriminative video representation with temporal order for micro-expression recognition," in *IEEE Int. Conf. Acoust. Speec. Sign. Proc.*, 2019, pp. 2112–2116.

[49] X. Hong, G. Zhao, S. Zafeiriou, M. Pantic, and M. Pietikäinen, "Capturing correlations of local features for image representation," *Neurocomputing*, vol. 184, pp. 99–106, 2016.

[50] J. A. Ruiz-Hernandez and M. Pietikäinen, "Encoding local binary patterns using the re-parametrization of the second order gaussian jet," in *IEEE Int. Conf. Works. Automat. FG Recog.*, 2013, pp. 1–6.

[51] S. K. A. Kamarol, M. H. Jaward, J. Parkkinen, and R. Parthiban, "Spatiotemporal feature extraction for facial expression recognition," *IET Image Processing*, vol. 10, no. 7, pp. 534–541, 2016.

[52] X. Huang, S.-J. Wang, G. Zhao, and M. Pietikainen, "Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection," in *IEEE Int. Conf. Comput. Vis. Works.*, 2015, pp. 1–9.

[53] X. Huang, S.-J. Wang, X. Liu, G. Zhao, X. Feng, and M. Pietikainen, "Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition," *IEEE Trans. Affect. Comput.*, vol. 10, pp. 32–47, 2017.

[54] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expression detection in hi-speed video based on facial action coding system (FACS)," *IEICE Trans. on Information and Systems*, vol. 96, no. 1, pp. 81–92, 2013.

[55] T. F. Cootes, G. J. Edwards, and C. J. Taylor, "Active appearance models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 23, no. 6, pp. 681–685, 2001.

[56] A. K. Davison, M. H. Yap, and C. Lansley, "Micro-facial movement detection using individualised baselines and histogram-based descriptors," in *2015 IEEE International Conference on Systems, Man, and Cybernetics*, 2015, pp. 1864–1869.

[57] M. Chen, H. T. Ma, J. Li, and H. Wang, "Emotion recognition using fixed length micro-expressions sequence and weighting method," in *IEEE Int. Conf. Comput. Real. Robot.*, 2016, pp. 427–430.

[58] Z. Lu, Z. Luo, H. Zheng, J. Chen, and W. Li, "A delaunay-based temporal coding model for micro-expression recognition," in *Asian Conference on Computer Vision*, 2014, pp. 698–711.

[59] S.-J. Wang, W.-J. Yan, G. Zhao, X. Fu, and C.-G. Zhou, "Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features," in *Workshop at the European Conference on Computer Vision*, 2014, pp. 325–338.

[60] G. Zhao and M. Pietikäinen, "Visual speaker identification with spatiotemporal directional features," in *International Conference Image Analysis and Recognition*. Springer, 2013, pp. 1–10.

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

[62] J. He, J.-F. Hu, X. Lu, and W.-S. Zheng, "Multi-task mid-level feature learning for micro-expression recognition," *Pattern Recognition*, vol. 66, pp. 44–52, 2017.

[63] Y.-H. Oh, A. C. Le Ngo, J. See, S.-T. Liong, R. C.-W. Phan, and H.-C. Ling, "Monogenic riesz wavelet representation for micro-expression recognition," in *IEEE Internationa conference on DSP*, 2015, pp. 1237–1241.

[64] Y.-H. Oh, A. C. Le Ngo, R. C.-W. Phari, J. See, and H.-C. Ling, "Intrinsic two-dimensional local structures for micro-expression recognition," in *IEEE Int. Conf. Acoust. Speec. Sign. Proc.*, 2016, pp. 1851–1855.

[65] O. Fleischmann, *2D signal analysis by generalized Hilbert transforms*, Thesis, University of Kiel, 2008.

[66] P. Zhang, X. Ben, R. Yan, C. Wu, and C. Guo, "Micro-expression recognition system," *Optik-International Journal for Light and Electron Optics*, vol. 127, no. 3, pp. 1395–1400, 2016.

[67] 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 Trans. Affect. Comput.*, vol. 9, no. 4, pp. 563–577, 2018.

[68] A. C. Le Ngo, Y.-H. Oh, R. C.-W. Phan, and J. See, "Eulerian emotion magnification for subtle expression recognition," in *IEEE Int. Conf. Acoust. Speec. Sign. Proc.*, 2016, pp. 1243–1247.

[69] S.-J. Wang, H.-L. Chen, W.-J. Yan, Y.-H. Chen, and X. Fu, "Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine," *Neural Processing Letters*, vol. 39, no. 1, pp. 25–43, 2014.

[70] X. Ben, P. Zhang, R. Yan, M. Yang, and G. Ge, "Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation," *Neural Computing and Applications*, vol. 27, no. 8, pp. 2629–2646, 2016.

[71] 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 Trans. Image Process.*, vol. 24, no. 12, pp. 6034–6047, 2015.

[72] D. Sun, S. Roth, and M. J. Black, "A quantitative analysis of current practices in optical flow estimation and the principles behind them," *International Journal of Computer Vision*, vol. 106, no. 2, pp. 115–137, 2014.

[73] M. Shreve, S. Godavarthy, V. Manohar, D. B. Goldgof, and S. Sarkar, "Towards macro-and micro-expression spotting in video using strain patterns." in *2009 Workshop on Applications of Computer Vision (WACV)*, 2009, pp. 1–6.

[74] M. Shreve, J. Brizzi, S. Fefilatyev, T. Luguev, D. Goldgof, and S. Sarkar, "Automatic expression spotting in videos," *Image & Vision Computing*, vol. 32, no. 8, pp. 476–486, 2014.

[75] D. Patel, G. Zhao, and M. Pietikäinen, "Spatiotemporal integration of optical flow vectors for micro-expression detection," in *International Conference on Advanced Concepts for Intelligent Vision Systems*, 2015, pp. 369–380.

[76] Y. Guo, B. Li, X. Ben, J. Zhang, R. Yan, and Y. Li, "A magnitude and angle combined optical flow fea-

[77] ture for micro-expression spotting," *IEEE Multimedia*, DOI(identifier)10.1109/MMUL.2021.3058017.

[77] S.-J. Wang, S. Wu, X. Qian, J. Li, and X. Fu, "A main directional maximal difference analysis for spotting facial movements from long-term videos," *Neurocomputing*, vol. 230, pp. 382–389, 2016.

[78] S. Polikovsky, Y. Kameda, and Y. Ohta, "Detection and measurement of facial micro-expression characteristics for psychological analysis," *Kameda's Publication*, vol. 110, pp. 57–64, 2010.

[79] A. Moilanen, G. Zhao, and M. Pietikäinen, "Spotting rapid facial movements from videos using appearance-based feature difference analysis," in *International Conference on Pattern Recognition (ICPR)*, 2014, pp. 1722–1727.

[80] Z. Xia, X. Feng, J. Peng, X. Peng, and G. Zhao, "Spontaneous micro-expression spotting via geometric deformation modeling," *Computer Vision and Image Understanding*, vol. 147, pp. 87–94, 2016.

[81] W.-J. Yan, S.-J. Wang, Y.-H. Chen, G. Zhao, and X. Fu, "Quantifying micro-expressions with constraint local model and local binary pattern," in *Workshop at the European Conference on Computer Vision*, 2014, pp. 296–305.

[82] 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 *2015 3rd IAPR Asian Conference on Pattern Recognition (ACPR)*, 2015, pp. 665–669.

[83] S.-T. Liong, J. See, R. C.-W. Phan, A. C. Le Ngo, Y.-H. Oh, and K. Wong, "Subtle expression recognition using optical strain weighted features," in *Asian Conference on Computer Vision*, 2014, pp. 644–657.

[84] F. Xu, J. Zhang, and J. Z. Wang, "Microexpression identification and categorization using a facial dynamics map," *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 254–267, 2017.

[85] 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 Trans. Affect. Comput.*, vol. 7, no. 4, pp. 299–310, 2016.

[86] Y.-J. Liu, B.-J. Li, and Y.-K. Lai, "Sparse MDMO: Learning a discriminative feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, 2019. [Online]. Available: https://doi.org/10.1109/TAFFC.2018.2854166

[87] R. Chaudhry, A. Ravichandran, G. Hager, and R. Vidal, "Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2009, pp. 1932–1939.

[88] G. Ras, M. van Gerven, and P. Haselager, *Explanation Methods in Deep Learning: Users, Values, Concerns and Challenges*. Cham: Springer International Publishing, 2018, pp. 19–36.

[89] M. Bartlett, G. Littlewort, J. Whitehill, E. Vural, T. Wu, K. Lee, A. Erçil, M. Cetin, and J. Movellan, "Insights on spontaneous facial expressions from automatic expression measurement," *Dynamic faces: Insights from Experiments and Computation*, pp. 211–238, MIT Press, 2010.

[90] U. Hess and R. E. Kleck, "Differentiating emotion elicited and deliberate emotional facial expressions," *European Journal of Social Psychology*, vol. 20, no. 5, pp. 369–385, 1990.

[91] Y. Guo, C. Xue, Y. Wang, and M. Yu, "Micro-expression recognition based on CBP-TOP feature with ELM," *Optik-International Journal for Light and Electron Optics*, vol. 126, no. 23, pp. 4446–4451, 2015.

[92] S. Zhang, B. Feng, Z. Chen, and X. Huang, "Micro-expression recognition by aggregating local spatio-temporal patterns," in *International Conference on Multimedia Modeling*, 2017, pp. 638–648.

[93] D. Patel, X. Hong, and G. Zhao, "Selective deep features for micro-expression recognition," in *International Conference on Pattern Recognition*, 2017, pp. 2258–2263.

[94] 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, pp. 1745–1756, 2017.

[95] X.-l. Hao and M. Tian, "Deep belief network based on double weber local descriptor in micro-expression recognition," in *Advanced Multimedia and Ubiquitous Engineering*, 2017, pp. 419–425.

[96] S.-J. Wang, B.-J. Li, Y.-J. Liu, W.-J. Yan, X. Ou, X. Huang, F. Xu, and X. Fu, "Micro-expression recognition with small sample size by transferring long-term convolutional neural network," *Neurocomputing*, vol. 312, pp. 251–262, 2018.

[97] H.-Q. Khor, J. See, R. C. W. Phan, and W. Lin, "Enriched long-term recurrent convolutional network for facial micro-expression

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/TPAMI.2021.3067464, IEEE Transactions on Pattern Analysis and Machine Intelligence

SUBMITTED TO IEEE T-PAMI

20

recognition," in *IEEE Int. Conf. Automat. FG Recog.*, 2018, pp. 667–674.

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

[99] X. Zhu, X. Ben, S. Liu, R. Yan, and W. Meng, "Coupled source domain targetized with updating tag vectors for micro-expression recognition," *Multimedia Tools & Applications*, vol. 77, no. 3, pp. 3105–3124, 2018.

[100] 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.

[101] Y. Zong, W. Zheng, Z. Cui, G. Zhao, and B. Hu, "Toward bridging microexpressions from different domains," *IEEE Transactions on Cybernetics*, 2019.

[102] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, 2013, pp. 3444–3451.

[103] S. Milborrow and F. Nicolls, "Active shape models with SIFT descriptors and mars," in *2014 International Conference on Computer Vision Theory and Applications (VISAPP)*, vol. 2, 2014, pp. 380–387.

[104] C. Tomasi and T. Kanade, "Detection and tracking of point features," School of Computer Science, Carnegie Mellon Univ. Pittsburgh,1991.

[105] Y. Han, B. Li, Y. Lai, and Y. Liu, "CFD: A collaborative feature difference method for spontaneous micro-expression spotting," in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 1942–1946.

[106] D. Cristinacce and T. F. Cootes, "Feature detection and tracking with constrained local models." in *British Machine Vision Conference*, vol. 1, no. 2, 2006, pp. 929–938.

[107] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2nd Edition)*. NY, USA: Wiley-Interscience, 2000.

[108] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *J. Mach. Learn. Res.*, vol. 15, no. 1, pp. 3133–3181, 2014.

[109] N. Van Quang, J. Chun, and T. Tokuyama, "Capsulenet for micro-expression recognition," in *IEEE Int. Conf. Automat. FG Recog.*, 2019, pp. 1–7.

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

[111] Z. Xia, X. Hong, X. Gao, X. Feng, and G. Zhao, "Spatiotemporal recurrent convolutional networks for recognizing spontaneous micro-expressions," *IEEE Trans. Multimedia*, 2019. [Online]. Available: https://doi.org/10.1109/TMM.2019.2931351

[112] V. Pérez-Rosas, M. Abouelenien, R. Mihalcea, and M. Burzo, "Deception detection using real-life trial data," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, 2015, pp. 59–66.

[113] D. Chen, S. Ren, Y. Wei, X. Cao, and J. Sun, "Joint cascade face detection and alignment," in *European Conference on Computer Vision*, 2014, pp. 109–122.

[114] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines." *ACM Trans. Intelligent Systems and Technology*, vol. 2, no. 3, pp. 1–27, 2011.

[115] T. Pfister, X. Li, G. Zhao, and M. Pietikainen, "Recognising spontaneous facial micro-expressions," in *Proceedings of the 2011 International Conference on Computer Vision*, 2011, pp. 1449 – 1456.

[116] J. Wright, A. Ganesh, S. Rao, Y. Peng, and Y. Ma, "Robust principal component analysis: Exact recovery of corrupted low-rank matrices via convex optimization," in *Advances in Neural Information Processing Systems*, 2009, pp. 2080–2088.

[117] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 770–778.

[118] 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 (ICPR)*. IEEE, 2018, pp. 946–951.

[119] D. H. Kim, W. J. Baddar, and Y. M. Ro, "Micro-expression recognition with expression-state constrained spatio-temporal feature representations," in *Proceedings of the 24th ACM International Conference on Multimedia*, 2016, pp. 382–386.

[120] Y. Li, X. Huang, and G. Zhao, "Can micro-expression be recognized based on single apex frame?" in *IEEE International Conference on Image Processing (ICIP)*, 2018, pp. 3094–3098.

[121] 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 & Gesture Recognition (FG 2018)*, 2018, pp. 657–661.

[122] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *Proc. IEEE Int. Conf. Comput. Vis. Pattern Recog. Workshops*, 2010, pp. 94–101.

[123] Q. Yang, Y. Liu, T. Chen, and Y. Tong, "Federated machine learning: Concept and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 10, no. 2, pp. 1–19, 2019.

[124] G. Warren, E. Schertler, and P. Bull, "Detecting deception from emotional and unemotional cues," *Journal of Nonverbal Behavior*, vol. 33, no. 1, pp. p.59–69, 2009.

[125] S. Li and W. Deng, "Deep facial expression recognition: A survey," *arXiv preprint arXiv:1804.08348*, 2018.

[126] C. Rudin, "Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead," *Nature Machine Intelligence*, vol. 1, no. 5, pp. 206–215, 2019.

Xianye Ben received the Ph.D. degree from the College of Automation, Harbin Engineering University, in 2010. She is now a Professor in the School of Information Science and Engineering, Shandong University, China. She received the Excellent Doctorial Dissertation awarded by Harbin Engineering University. She was also enrolled by the Young Scholars Program of Shandong University.

Yi Ren received the B.S. degree in electronic and information engineering from the School of Physics and Electronic Science, Shandong Normal University, Jinan, China, in 2016. Her current research interests include micro-expression spotting/detection and recognition.

Junping Zhang (M'05) is a professor at School of Computer Science, Fudan University since 2011. His research interests include machine learning, image processing, biometric authentication, and intelligent transportation systems. He has been an associate editor of the IEEE INTELLIGENT SYSTEMS since 2009.

Su-Jing Wang (SM'19) received the Ph.D. degree from the College of Computer Science and Technology of Jilin University in 2012. He is now an Associate Researcher in Institute of Psychology, Chinese Academy of Sciences. He was named as *Chinese Hawkin* by the Xinhua News Agency. His current research interests include pattern recognition, computer vision and machine learning. He serves as an associate editor of Neurocomputing (Elsevier).

Kidiyo Kpalma joined the Institut National des Sciences Appliquees de Rennes (INSA), and received the PhD degree in 1992. In 1994, he became Associate Professor at INSA where he teaches Analog/Digital Signal processing an Automatic. After his HDR (Habilitation a diriger des recherches) degree from the University of Rennes 1 in 2009, he obtain the position of Professor with INSA since 2014. As a member of IETR UMR CNRS 6164, his research interests include pattern recognition, semantic image segmentation, facial micro-expression and saliency object detection.

Weixiao Meng (SM'10) received the B.Eng., M. Eng., and Ph.D. degrees from Harbin Institute of Technology (HIT), Harbin, China, in 1990, 1995, and 2000, respectively. He is a professor of the School of Electronics and Information Engineering of HIT. He is the Chair of IEEE Communications Society Harbin Chapter and a Fellow of the China Institute of Electronics. He won Chapter of the Year Award, Asia Pacific Region Chapter Achievement Award and Member & Global Activities Contribution Award in 2018.

Yong-Jin Liu (SM'16) is a tenured full professor with Department of Computer Science and Technology, Tsinghua University, China. He received his B.Eng degree from Tianjin University, China, in 1998, and the PhD degree from the Hong Kong University of Science and Technology, Hong Kong, China, in 2004. His research interests include cognition computation and computer vision. See more information at https://cg.cs.tsinghua.edu.cn/people/~Yongjin/Yongjin.htm