The Visual Computer (2020) 36:445–468
https://doi.org/10.1007/s00371-018-1607-6

SURVEY

Micro-expression recognition: an updated review of current trends, challenges and solutions

Kam Meng Goh[1] · Chee How Ng[1] · Li Li Lim[1] · U. U. Sheikh[2]

Published online: 27 October 2018

Abstract

Micro-expression (ME) recognition has attracted numerous interests within the computer vision circle in different contexts particularly, localization, magnification, and recognition. Challenges in these areas remain relevant due to the nature of ME's split-second transition with minute intensity levels. In this paper, a comprehensive state-of-the-art analysis of ME recognition and detection challenges are provided. Contemporary solutions are categorized into low-level, mid-level, and high-level solutions with a review of their characteristics and performances. This paper also provides possible extensions to basic methods, highlight, and predict emerging trends. A thorough analysis of mainstream ME datasets is also provided by elucidating each of their advantages and limitations. This survey gives readers an understanding of ME recognition and an appreciation of future research direction in ME recognition systems.

Keywords Classification · Dataset · Feature extraction · Micro-expression · Pre-processing · Spotting

1 Introduction

Micro-expression (ME) is a transitory motion of the human face that usually lasts between 1/25 and 1/5 s [1]. The study of ME provides the ability to expose genuine emotions that occur briefly and unintentionally even when true emotions are deliberately masked. Ongoing research indicates that ME usually occurs in high stake situations and is challenging to detect in real time [2]. Contrasting from normal facial expressions, it is difficult to intentionally produce or neutralize ME which makes for effective evidence of lie detection that happens in scenarios where a person has something to lose or gain [3].

Ekman has developed the Micro-Expression Training Tool (METT) [4] to train people on ME identification. Accordingly, Ekman has categorized human emotions into seven universal emotions, i.e. anger, happiness, sadness, disgust, surprise, fear, and contempt [5–8]. Additionally, Ekman and Friesen have also introduced Facial Action Coding System (FACS) to define facial expressions through action units (AUs) [9]. AU is an observable component of facial movement where distinct facial areas are used to detect fine-grained expression changes on faces [10]. There are currently a total of 44 AUs that happen independently or simultaneously with other AUs to express an emotion. However, the accuracy of human recognition with AU is only about 40% [11] due to short-lived ME occurrences. The maximum accepted time duration for a ME to occur is 0.5 *s* at low intensity [12]. Hence, the demand for a system to identify, recognize, and analyse ME is essential [13].

A vast range of research has been done on ME recognition using computer vision solutions. However, current challenges such as environmental variation, spontaneous subtle motion, and imbalanced datasets that greatly impact detection and recognition accuracy are yet to be solved and detailed as follows:

✉ Kam Meng Goh
gohkm@tarc.edu.my

Chee How Ng
C.How_93@hotmail.com

Li Li Lim
lllim@tarc.edu.my

U. U. Sheikh
usman@fke.utm.my

[1] Department of Electrical and Electronics Engineering, Faculty of Engineering and Technology, Tunku Abdul Rahman University College, Jalan Genting Kelang, 53300 Wilayah Persekutuan Kuala Lumpur, Malaysia

[2] Faculty of Electrical Engineering, Universiti Teknologi Malaysia, 81310 Skudai, Johor, Malaysia

- *Environmental factor*
  Environmental variation is the most challenging issue in ME recognition which includes illumination variation [13] and head-pose variation [5]. In illumination variation, most features are heavily dependent on the intensity of pixels change, such as Local Binary Pattern (LBP) [14], optical flow [15, 16], or Histogram of Oriented Gradient (HOG) [5]. Lighting variation may cause incorrect feature estimation [13], head movement, or head-pose variation and be misinterpreted as ME. A subtle head movement significantly impacts face component change, hence affecting detection accuracy [5, 17].
- *Spontaneous and subtle motion of facial movement*
  The low intensity of subtle and spontaneous facial movement is a major challenge for ME recognition [5] which renders emotion recognition non-distinguishable through the naked eyes [12, 18–20]. Often, the classifier may also incorrectly interpret motions as a neutral face [2]. Therefore, methods to magnify [5, 21, 22] and enhance subtle emotion at the pre-processing stage are crucial.
- *Imbalanced dataset*
  There are a few publicly available datasets that focus on ME recognition, such as Spontaneous Micro-expression Corpus (SMIC) [8], Chinese Academy of Sciences Micro-Expression (CASME) [6], Chinese Academy of Sciences Micro-Expression 2 (CASME II) [7], and Chinese Academy of Sciences Macro-Expressions and Micro-Expressions CAS(ME)^2 [17]. Although recommended in evaluating ME recognition system, their imbalanced data distribution across expressions may lead to biases in results [23]. Furthermore, samples of available datasets are usually taken in controlled environment with even illumination and rigid positions. Hence, well-tested algorithms using these datasets may not be suitable in normal conditions and provide for a demand for dataset in natural settings.

Numerous works were done to mitigate these challenges. In this paper, we provide the most comprehensive and in-depth study of existing solutions to these challenges and their respective advantages and limitations. Unlike related reviews [2], our contributions in this paper are as follows:

- A review of spotting methods is provided.
- All existing features are categorized into low-level, mid-level, and high-level features. A discussion of each feature characteristics in addressing ME challenges is also included.
- We analyse the evaluation procedures used to measure the performance of ME recognition, and also the performance of every method.
- We uncover the advantages and limitations of current ME datasets.

The overall contribution objective is to facilitate an understanding of current challenges, trends, and developments in this domain to ease efforts in enhancing ME recognition. This paper is organized in the following manner: Sect. 2 explains the context in ME recognition, while in Sect. 3, mainstream ME features in low-level, mid-level, and high-level representation used for recognition are discussed. Section 4 identifies a review of ME spotting, and recognition results and discussion are presented in Sect. 5. We conclude our review in the final section.

2 Context

2.1 Dataset

In all ME datasets, participants are requested to be dispassionate or more specifically, required to keep a “poker face” during experiments and to report their emotions for each stimulant post-recording.

The Polikovsky’s Database [24], USF-HD [25], and York Deception Detection Test (York-DDT) [26] are the first datasets used for ME recognition testing. However, these datasets are unpopular due to their inadequacies. For example, Polikovsky’s Database [24] and USF-HD [25] data are collected by asking participants to intentionally pose or mimic emotions which contradict the spontaneous nature of ME. Meanwhile, York-DDT [26] contains only 18 spontaneous ME samples which are insufficient for data analysis. In this section, we focus only on the advantages and limitations of all publicly available spontaneous expressions datasets. Considering that MEs usually happen spontaneously and involuntarily, the dataset should consist of expressions of such nature. State-of-the-art approaches are commonly tested in popular datasets, such as SMIC [8], CASME [6], CASME II [7], and CAS(ME)^2 [17]. On the other hand, Spontaneous Actions and Micro-Movements (SAMM) dataset was introduced [27] to focus on AU recognition based on 13 different ethnicities of 32 participants. These datasets focus on micro-movement recognition instead of emotion identification. The details of these datasets are displayed in Table 1.

The SMIC consists of 16 valid subjects with 164 samples captured at 100 frames per second in a controlled environment which is the extended version of the original SMIC dataset consisting only 77 samples with emotions divided into three categories, namely positive, negative, and surprise. Positive class consists of happy emotion with 51 samples, while the negative class consists of 70 samples from four emotions which are sadness, anger, fear, and disgust. The remaining 43 samples are categorized to surprise. All micro-emotions in this dataset were induced by watching video clips with a maximum of 4.3 min length. However, the dataset is not labelled with AU, and the index of the most expressive

Table 1 Summary of spontaneous ME datasets

	SMIC HS/VIS/NIR	CASME	CASME II	CAS (ME)^2	SAMM
Number of Participants	16/8/8	35	35	22	32
Number of Samples	164/71/71	195	247	57	159
Number of Emotion	3	7	5	4	7
Resolution	640 × 480	640 × 480/720 × 1280	640 × 480	640 × 480	2040 × 1088
Frame rate	100/25/25	60	200	30	200
FACS coded	No	Yes	Yes	Yes	Yes
Ethnicities	3	1	1	1	13

frames (apex) remains unknown. On the other hand, SMIC also contains samples recorded from 8 subjects using normal speed cameras (VIS) and infrared cameras (NIR). For each type of cameras, 71 samples containing ME are provided, respectively. Both VIR and NIR datasets in SMIC are recorded with 25 frames per second.

CASME consists of 195 samples out of 1500 facial expressions from 19 valid subjects at 60 frames per second. It consists of 8 emotion classes which are happiness (5 samples), sadness (6 samples), disgust (88 samples), surprise (20 samples), contempt (3 samples), fear (2 samples), repression (40 samples), and tense (28 samples), where sample distribution among classes is highly imbalanced. These emotions are AU labelled with face size of 150 × 190 pixels. Similar to SMIC, videos were shown to participants to induce ME, and participants were requested to suppress all emotional expression on their face; otherwise, the amount of token given will be deducted. The duration of the displayed videos was about 1–4 min, and collected data were analysed with AU identification from the videos.

CASME II is an improved dataset version of CASME where the samples are increased to 247 with 26 valid subjects. Intensive selection was done to pick the best samples out of 2500 facial expressions. A few videos with a maximum duration of 3 min were displayed to trigger participants' emotion. In addition, recording is done at 200 frames per second to collect facial expression in a controlled environment. After recording, participants were required to report their feelings for tailoring purposes. Expressions are classified into 5 emotion classes which are happiness (33 samples), disgust (60 samples), surprise (25 samples), repression (27 samples), and others (102 samples) with face sizes cropped to 280 pixels × 340 pixels. CASME II also resolved illumination problems present in the previous dataset by providing a steady and high-intensity lighting environment. Moreover, it has greater temporal and spatial resolution than CASME as well as larger sample size. One drawback of CASME II is the imbalanced sample distributions among classes. Also, the participants were limited to youths within an ethnicity. Figure 1 shows one of the frame sequences of disgust in CASME II.

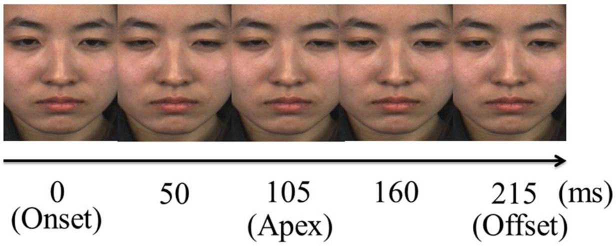


Fig. 1 A demonstration of frame sequence from CASME II. The brows have been lowered and drawn together, while lower eyelids are tightened in apex frame, which indicates disgust

The frame that an emotion begins indicates the onset frame, while the apex frame contains the most expressive emotion. Offset frame indicates the frame where emotion is back to neutral. From Fig. 1, the difference between the apex frame and onset frame is insignificant. To show AU changes in apex frame, we illustrate examples from CASME II in Fig. 2.

Figure 2 shows that distinct ME can be represented using different AU. Happiness usually occurs in AU12, where lip corners are pulled obliquely, while inner brows are raised (AU1) when sadness occurs. AU4 and AU10 happen together for disgust, where the eyebrows are lowered with raised lips.

It was reported in Qu et al. [17] that older datasets are unsuitable for ME spotting due to short duration samples. Hence, the CAS(ME)^2 dataset that consists of macro- and micro-expressions in longer duration was proposed. The dataset is divided into two portions for expressions; part A with 87 long videos with both types of expression; and part B with 300 cropped spontaneous macro-expression along with 57 samples of spontaneous ME from 22 participants. Expres-

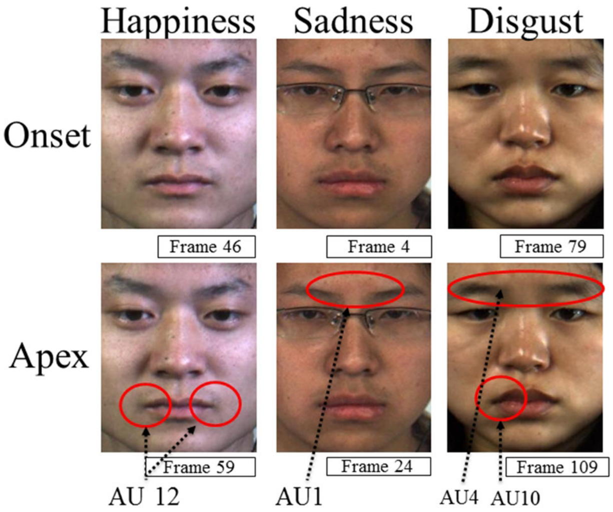


Fig. 2 Different emotions revealed by observing the action units at apex frame from CASME II datasets

sions in this dataset are classified into four classes which are positive (8 samples), negative (21 samples), surprise (9 samples), and others (19 samples). Similar to SMIC, a positive class is related to happiness, amusement, and delight, while a negative class constitutes the opposite, e.g. disgust, anger, and fear with AUs labelled in each sample. However, samples were collected with a relatively low frame rate of 30 frames per second in relatively small number of samples compared to the previous dataset. This affects the accuracy of classification and is unsuitable for high-level approaches.

According to Davison et al. [27] ME datasets to date lack ethnic diversity. Hence, Spontaneous Actions and Micro-Movements (SAMM) dataset was introduced [27]. In this dataset, participants from 13 different ethnicities were recorded. Facial expressions were taken at 200 frames per second in controlled lighting condition to prevent flickering. Unlike previous dataset collection, participants were first required to fill a questionnaire before they proceed to experiments. Based on the answers given, the experiment conductor displayed videos that are relevant to the answers. For example, if a participant mentioned the fear of heights, a video of a bungee jump was displayed to induce fear in the targeted participant. All recorded videos are FACS coded with less emphasis on emotional labelling.

Besides the aforementioned datasets, weak expressions in macro-expressions datasets are also used for ME recognition. Yu et al. [28] used the CK+ macro-expressions dataset as their benchmarks. These datasets contain basic 6 emotions, with frames of strong and weak expressions. Instead of using micro-expressions dataset, the authors used weak expressions for their work. Since this dataset is not commonly used in ME recognition, it is omitted from Table 1 and further details of their proposed work are discussed in Sect. 3.3.

2.2 General pipeline

The ME recognition process can be divided into image acquisition, face detection, pre-processing, ME spotting, feature extraction, and ME classification as illustrated in Fig. 3.

The use of high speed camera is essential to capture rapid and subtle motion of facial components which are undetectable through the naked eyes or low frame rate videos [29]. The facial images are pre-processed before undergoing ME spotting and recognition. First, a face will be detected post-image acquisition and subsequently segmented from the background followed by prototypical face registration [30] to counter head-pose variation.

With a detected face, several pre-processing steps are implemented to overcome lighting variation [31] or noise attack [32]. Subsequent subtle motion magnification work eases the recognition process [21], as detailed in Sect. 2.3.2. The enhanced images are then delivered to ME spotting, feature extraction, and classification. However, work focusing on ME spotting is less. It is common to bypass the spotting stage and focus on features selection and classification methods. Thus, research on ME spotting from images or videos is a potential future research direction.

2.3 Pre-processing

Pre-processing in ME recognition usually involves face detection, face registration, motion magnification, and temporal normalization. After normalization of detected faces in face registration, noises are filtered and available features are strengthened for better performance. Methods for processing usually involve magnification of subtle features [21] and temporal normalization [33], while noise removal is done using a Wiener [32] or Gaussian filter [34].

2.3.1 Face detection and registration

Viola and Jones [35] pioneered the introduction of the first framework in 2001 in this domain. The Viola–Jones cascade classifier is a machine learning approach for real-time, robust object detection that detects faces in an image.

Multi-task learning was also proposed for facial detection based on a mixture of trees model with shared pool of parts [36]. On the other hand, a joint cascade method was proposed in [37] for simultaneous facial and landmark points detection. In deep learning, Matsugu et al. [38] spearheaded a rule-based algorithm in facial expression recognition with convolutional neural network for face detection. The proposed work claims to be robust to translation, scale, and pose. Additionally, various CNN-based face detection methods have been proposed to overcome pose and illumination variation, such as NPDFace [39] and others [40–42]. Recently, Ranjan et al. [43] proposed a single convolutional neural net-

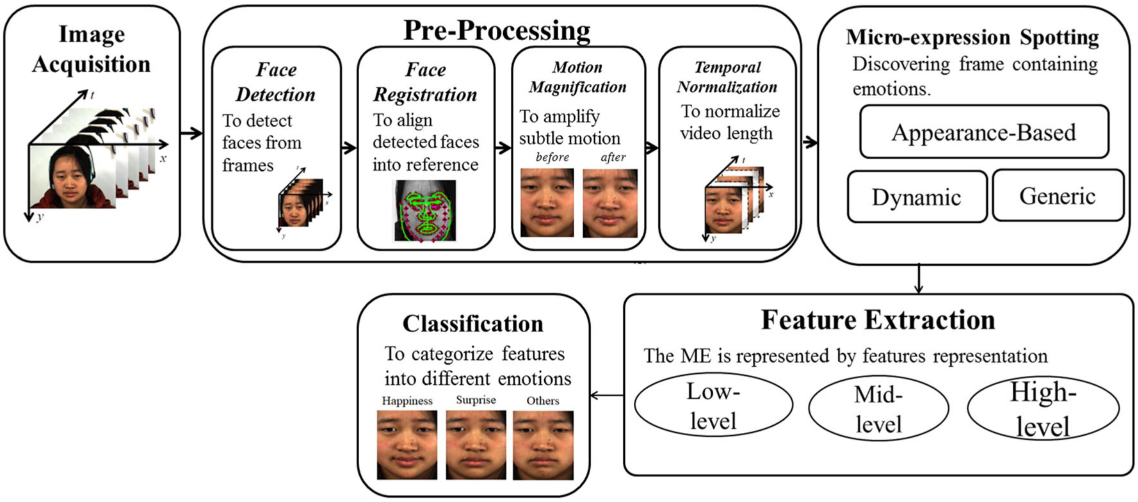


Fig. 3 General pipeline of ME recognition process

work (CNN) model to classify an image as a face or non-face. In this work, a selective search algorithm in region-CNN is used to generate region proposals for face search in an image. The method was tested in Annotated Face in-the-Wild (AFW) [36] dataset and PASCAL faces dataset [44], with remarkable results. However, this approach does not localize well with the actual face region.

The face registration stage aligns a detected face onto a reference face [31]. This is an important step to enhance performance and to handle varying head-pose issues. To date, considerable work has been done on face registration which can be categorized as fiducial landmark points and generic approaches [30].

In landmark based approaches, it is common to detect landmark points on a face and then transform the input to a prototypical face. In this domain, active shape model (ASM) [45] is widely used in modelling facial features [7, 8, 46–54] due to its flexibility and reliability. ASM applies 68 landmark points that are repeatedly revised to refine approximations of pose, scale, and shape of model image objects. Studies show that increased landmarks points are beneficial to performance as there is lesser registration error and is more resilient to head-pose variation [30]. Sariyanidi et al. [31] reported that some landmark localizers are illumination invariant [55–57], but these methods are sensitive to jitter among consecutive frames due to localization error. Recently, OpenPose library is introduced to detect body, hand, and facial landmark in real time [58]. In this library, 70 key points are used to perform face registration where facial points detection is done based on procedures in [59]. In recent years, CNN-based landmark localization methods [60, 61] have been proposed which outperform other methods. In [43], the authors used 21 fiducial points in their CNN and applied landmarks-based non-maximum suppression (L-NMS) to improve performance.

Generic techniques for face registration, such as fractional Fourier transform [62], Fourier-Mellin transform [63], or Radon transform [64], are sensitive to lighting changes. The Lucas–Kanade (LK) method is proposed for face registration where minimization between two frames is done by assuming changes of pixels between two neighbouring frames which is small [65]. Several extensions of LK with optimization have been adopted [66–68], but the pixel-based method remains sensitive to illumination changes. Therefore, the Gabor filter is implemented to overcome this issue [67].

2.3.2 Motion magnification

ME facial movements are persistently hard to distinguish due to its subtleness. Hence, motion magnification techniques [5, 21, 22] are introduced to increase distinguishing powers between different motions. The Eulerian video magnification (EVM) method is commonly used to magnify subtle motions by enhancing motion differences. EVM magnifies both motions and colours to ease feature extraction process. However, there is still a lack of magnification work to date.

2.3.3 Temporal normalization

Temporal interpolation model (TIM) method is commonly used to normalize video lengths [46, 47, 54, 69]. It is specifically designed to tackle spontaneous subtle expressions, which are unexpected and difficult to detect accurately. TIM is also used to interpolate fewer frames to remove redundant faces without emotions. However, this method does not improve the effectiveness of recognition performance [33]. Instead, Xu et al. [13] used linear interpolation as it is capable of indicating motion patterns with less error. Meanwhile, Le Ngo et al. [33] claimed that TIM partially removes redundant information at regularly separated positions without knowledge of sparse information in the frame which might be accidentally removed. Thus, Sparsity-promoting dynamic mode decomposition (DMDSP) [33] was proposed to overcome the drawbacks of TIM. The demonstrated results show that DMDSP provides better result compared to TIM.

2.4 Classification

Classification normally refers to the categorization of emotions based on selected features input. Training and testing are normally involved in classification, where the training process teaches the classifier to recognize ME based on features and annotation given, and testing checks the accuracy of the classifier. There are various supervised crisp classification methods that have been proposed for ME recognition, such as Support Vector Machine (SVM) [70], Multiple Kernel Learning (MKL) [46–48], *k*-Nearest Neighbour (*k*NN) [71–73], Random Forest [46], Extreme Learning Machine (ELM) [74–76], Linear Discriminant Analysis (LDA) [72, 77], A Library for Large Linear Classification (LIBLINEAR) [78], and Softmax in CNN [79, 80]. Notably, instead of classifying ME, Polikovsky et al. [24] employed the *k*-mean cluster which is an unsupervised learning method for AU categorization by classifying facial cubes such as forehead, left and right eyebrows, left and right eyes, between the eyes, lower nose, mouth, left and right mouth corners, and chin. Subsequently, facial cubes are manually categorized into an AU.

Within high-level approach classification in ME recognition; fully connected layer [80, 81] in CNN is used. The conventional multilayer perceptron is used in fully connected layer to classify high-level feature extracted from convolution and max pooling layer into several classes with Softmax initiation. The fully connected layer is often referred as the last layer of CNN. Since this layer is Softmax initiated, it is termed as the Softmax classifier.

In lieu of the crisp classifier, Lim and Goh [82] recently addressed ME recognition as non-mutually exclusive cases [83]. They presumed that there are similar subtle motions crossover of different expressions, and this non-distinguish motions will bring ambiguity in feature modelling and deteriorate the recognition accuracy generated by conventional Binary/Crisp classifiers. To overcome this problem, Lim and Goh [82] utilize Fuzzy Qualitative Rank Classifier (FQRC) [83] to model features corresponding to the ME in the form of fuzzy tuple without constraint on feature extraction method, [84, 85]. This approach relaxes ambiguities by considering all possibilities in the processing pipeline and generates multi-class classification result by ranking (confidence value of the micro-expression) the outcomes.

3 Features for ME representation

Features are defined as representation of ME that is used for classification purpose with diverse features. Conventionally, face representations can be classified to spatial/spatial–temporal classes [2]. Due to rapid advancement and attention in high-level approach, coupled with large volume of reported works in low-level representation, this paper categorizes each approach into low-level, mid-level, and high-level representation in distinct sections.

3.1 Low-level representation

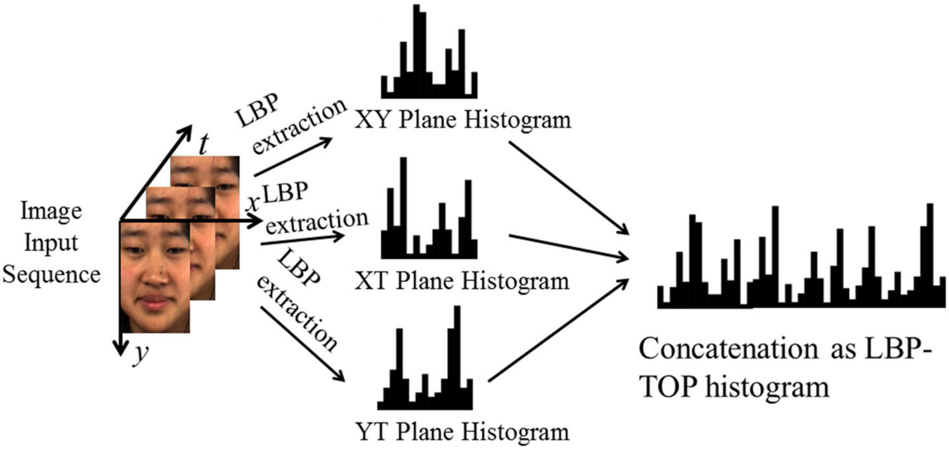
Due to the nature of publically available datasets with limited sizes, low-level representations are widely researched and emphasized in ME recognition. The features in this category are normally extracted from minor details of images, i.e. intensities [7], change of intensities in temporal [16], gradient [5], etc. Low-level features are normally represented in the form of descriptors containing a bunch of visual data cue without explicit semantic meaning/knowledge.

The extracted information is classified with basic classifiers, as mentioned in Sect. 2.4. In this paper, we briefly describe features in the following family: local binary pattern (LBP), optical flow, gradient based, and their respective variants.

3.1.1 Local binary pattern and its variants

LBP is greatly utilized in many applications especially face expression recognition due to its effective ability in face features extraction. Ojala et al. [14] are one of the first to introduce the LBP operator which is useful in describing textures. The LBP has a good tolerance when it comes to monotonic illumination change, but handles only spatial information. In order to extract ME features with low intensity value, temporal information is needed. Hence, Local Binary Pattern on Three-Orthogonal Planes (LBP-TOP) [86] is extended from the basic LBP for simultaneous spatial and temporal information extraction. LBP-TOP was first designed to extract macro-emotion in [86] and is further implemented in ME recognition in [7, 46]. A dynamic texture with motion and appearance are seen in a set of volume of three directions labelled *X*, *Y*, and *T*. The first two directions are spatial coordinates, while the third is time, also known as frame index. Based on *X*, *Y*, and *T* directions, *XY*, *XT*, and *YT* planes are computed. Every pixel of each plane will undergo extraction by the LBP operator, yielding its corresponding histogram. In order to obtain the feature vector, the three respective histograms are combined into one histogram. Yan et al. [7], Li et al. [8], Pfister et al. [46], Guo et al. [71], Adegun and Vadapalli [75], and House and Meyer [87] implement LBP-TOP features extraction with different facial detection and classi-

Fig. 4 LBP-TOP feature extraction on three different planes and its resultant histogram being concatenated



fication method for ME recognition. Figure 4 illustrates the process of LBP-TOP feature extraction.

Although LBP-TOP is computationally expensive [88], efforts were made to reduce computational complexity to improve the performance of LBP-TOP. Local Binary Pattern with Six Intersection Points (LBP-SIP) was introduced by means of LBP-TOP modification. LBP-SIP observes the neighbouring points of the centre pixel for the three-orthogonal planes [88], while LBP-SIP differs from the LBP-TOP approach where it takes each neighbouring point into account, considering only six unique intersection points (*A*, *B*, *D*, *E*, *F*, and *G*) that are located on the overlapping lines of the three-orthogonal planes [3].

These six unique intersection points will be used to form binary patterns with the spatiotemporal textures described at the centre point. Similar to LBP-TOP, it views the neighbouring points on two-dimensional planes as separate and concatenates them into one histogram descriptor. In terms of dimensionality, LBP-SIP produces 2.4 times shorter histogram length than LBP-TOP. Also, LBP-TOP's accuracy is lower compared to LBP-SIP on the SMIC and CASME II dataset.

The same authors also proposed Local Binary Pattern with Mean Orthogonal Planes (LBP-MOP) [32]. Instead of utilizing every video frame, a reduction in the computational complexity using mean planes for feature calculation is employed. LBP-MOP calculates the mean from three planes and computes its corresponding LBP for each plane stack. In short, this method compresses the spatial and temporal information before it begins to extract LBP features. In terms of feature extraction time, LBP-MOP takes 17.81 s shorter than LBP-TOP [32]. LBP-MOP also outperforms LBP-TOP in terms of accuracy when the leave-one-subject-out (LOSO) approach is used for validation.

Meanwhile, Centralized Binary Pattern from Three-Orthogonal Planes (CBP-TOP) is an enhanced version of LBP-TOP [76] that uses similar method as LBP-TOP for ME extraction. The length of the histogram is significantly reduced compared to the conventional LBP-TOP. CBP codes are obtained by comparing the pixel located at the centre with pairs of connecting neighbour points only. By placing the highest weight to the centre pixel, it emphasizes the importance of the centre pixel which improves discrimination with lower histogram length in CBP compared to the LBP. However, one disadvantage of CBP-TOP is that the computation focuses mostly on the centre pixel which requires longer computation duration with trade off in terms of accuracy as CBP-TOP outperforms LBP-TOP after testing in CASME dataset.

A new approach in encoding the LBP was proposed by Ruiz-Hernandez and Pietikäinen [48] with re-parametrization of the second local order Gaussian Jet to generate a more robust and consistent histogram for ME recognition. In this approach, second local order Gaussian Jets are first computed from video frames using Gaussian derivative. The Gaussian Jets are re-parametrized into angular coordinate (parameters) which are robust to illumination, rotation, or scaling changes. LBP-TOP is encoded using these angular parameters instead of the conventional intensity-based histogram to improve statistical stability and discriminative power. The criteria are set with the use of gradient orientation and second-order Gaussian derivatives. With information obtained through re-parametrization, the recognition process can be drastically enhanced by addressing unstable statistic in LBP-TOP. A benefit of re-parametrization is to facilitate the ability to describe complicated shapes that increase recognition accuracy. Moreover, it is statistically more stable compared to LBP-TOP for ME recognition despite producing a lot of redundant information due to its high dimensional feature.

LBP-TOP can also be integrated with integral projection to enhance discriminative factor. Huang et al. [54] proposed a combination of LBP-TOP with integral projection (IP) as an image projection that could provide supplementary shape information to enhance texture descriptor by obtaining horizontal and vertical projection using IP on difference images. The 1DLBP operator is then used to obtain features on the horizontal and vertical projections to obtain spatial information in the form of a histogram [54]. The 2DLBP is implemented on the motion texture image to obtain motion histogram. The vertical histogram, horizontal histogram, and motion histogram are then concatenated to form the final feature vector. Using this method, the shape information of ME can be maintained while improving discrimination information. The proposed method was tested on CASME II and SMIC dataset and outperforms the accuracy of conventional LBP-TOP and LBP-SIP. STLBP-IP, however, uses onset, apex, and offset frames to extract feature assuming that the onset frame is always a neutral emotion.

In order to overcome the drawbacks of STLBP-IP, Huang et al. [52] introduced a spatiotemporal feature descriptor which considers the shape attributes of dynamic texture information. In contrast, STLBP-IP with Revisited Integral Projection (STLBP-RIP) was introduced by breaking the premise that the onset frame is always a neutral face. STLBP-RIP integrates shape attributes with LBP and considers discriminative information to yield better results in ME recognition and feature selection in STLBP-RIP based on Laplacian method. The authors further extended their work to the discriminative version, STLBP-IP (DISTLBP-IP) by only selecting a few group of feature that brings the strongest discriminative power. Several experiments were conducted in [52] where three datasets namely CASME, CASME II and SMIC were used. STLBP-RIP outperforms the baseline and older feature extractors such as LBP-TOP, LBP-SIP and STCLQP in CASME. However, methods like Deep Convolutional Neural network (DCNN) on Compute Unified Device Architecture (CUDA) [89] and Local Spatiotemporal Directional Features (LSTD) [47] yield slightly improved performance. This is nonetheless questionable due to varying experimental setup. Overall, STLBP-RIP and DISTLBP-RIP may be able to achieve a reasonable and competitive accuracy, but the computation process is rather simple compared to other methods such as STCLQP and DCNN on CUDA.

Conventionally, LBP-TOP and its variants are applied on grey images. Interestingly, useful information can be provided by colour for face recognition which mirrors the characteristic of human insight [90]. Wang et al. [50] proposed a novel method where feature extraction method such as LBP-TOP is used on Tensor Independent Colour Space (TICS) rather than grey video. Other than TICS colour space, other colour spaces such as CIELab and CIELuv are introduced in [49]. From the experiments conducted in [50], the use of LBP-TOP on TICS is superior to grey and RGB space. The tensor is made up of four-dimensional arrays with the first and second consisting of spatial data. The third dimension array consists of temporal data, while the fourth dimension comprises the colour data. Colour information consists of three components, namely red (R), blue (B), and green (G) in RGB space and are transformed to TICS. The ME recognition performance is enhanced through the use of colour space model, where dynamic texture feature of ME is extracted from colour components. CIELab and CIELuv are used to replace TICS for accuracy testing. Overall, it is concluded that the use of colour component increases the recognition rate of ME.

Although LBP and its variants are widely used in ME recognitions, this family of features is normally applied on holistic faces, which causes massive redundancy in the feature histogram. Optimization method such as Robust Principal Component Analysis (RPCA) [52] can be used to reduce unnecessary data and subsequently increase recognition accuracy. In another approach, Lim and Goh [82] divide the face into several regions, followed by STLBP-IP implementation and fuzzy logic to identify areas of the face that highlights significant feature for recognition that shows promising results compared to the STLBP-IP.

Overall, the LBP method is a popular technique to extract low-level features. However, there are limitations; unlike dynamic feature, users are unable to identify the mechanism of recognition intuitively since the motion of facial component or action units are possibly unknown based on LBP features [13].

3.1.2 Optical flow and its variants

In contrast to the LBP family, the purpose of optical flow is to observe and capture non-rigid movement of facial component [72]. A robust optical flow method [15, 16] has been widely used in facial detection recognition through computation of displacement using brightness conservation principle. The aim of optical flow method is to estimate the strain magnitude. There are three phases of expression, i.e. start, peak and stop, where peaks of the expression can be translated into strain magnitude for subtle expressions which are initially difficult to find.

A derivative of optical flow as features was proposed by Liong et al. [91]. Similar to optical flow, it calculates subtle changes in facial motion. Optical strain differentiates relative deformation quantities of facial tissue or an object. Utilizing a displacement vector, it is possible to detect an object that is malleable in two-dimensional spaces. Liong et al. extended the work in [34, 91] with substantial improvements in [92] by introducing an optical strain map (OSM) computed from the magnitude of optical strain feature (OSF), optical strain weight (OSW), and the hybrid version of these two features. OSM also offers graphical illustration of motion concentration for each pixel in a video. In terms of spatial displacement, OSM is used to identify the region of an image frame with highest or lowest projecting motion. In contrast to the baseline LBP-TOP method, the optical strain method in [91] reported a higher accuracy in SMIC dataset. The newer optical strain experiment conducted in CASME II dataset also outperforms the baseline LBP-TOP method. Liong et al. [92] combined optical strain weight with optical strain features to achieve greater performance compared to LBP-TOP, LBP-SIP, STLBP-IP, and colour space in both SMIC and CASME II datasets. In addition, histogram bins that contribute to noise are discarded to improve recognition performance [93].

Since the recognition of ME is highly dependent on facial component motion, an extension of histogram of oriented optical flow (HOOF) was proposed [72, 94]. HOOF was originally proposed for human action recognition due to its scale-invariant and directional independence [94]. However, the HOOF histogram is easily affected by illumination changes. Therefore, Happy and Routray [72] introduced fuzzy histogram of oriented optical flow (FHOOF) to overcome HOOF's disadvantage. FHOOF constructs the histogram angular bin by collecting motion directions based on fuzzy membership functions where the Gaussian membership function is used in fuzzification process. Nonetheless, the computational process of finding membership values is tedious. To simplify, many fine histograms are formed and their membership values are set to coarse histogram. Fuzzy membership functions are also utilized to map fine histogram bins to coarse histogram bins where the authors additionally claim that FHOOF is insensitive to illumination changes.

Although FHOOF is illumination invariant, the weight assigned to FHOOF is highly dependent on motion magnitudes of various MEs, which differs from sample to sample. To overcome this drawback, the authors also introduced Fuzzy Histogram of Optical Flow Orientations (FHOFO) [72]. FHOFO assumes that negligible movement with insignificant magnitudes is discarded during histogram construction. Similar to FHOOF, utilizing the fuzzy membership function, the collected motion directions are assigned to various histogram bins. Furthermore, each sample may have disparate motion magnitude due to different magnitudes in ME interval. The recognition accuracy improves when motion magnitude is ignored. FHOOF and FHOFO achieved remarkable results compared to HOOF and LBP-TOP, as reported in [72]. Recently, Liong et al. [95] claim that a single apex frame is sufficient to identify hidden emotion, and extended HOOF into Bi-weighted oriented optical flow (Bi-WOOF) as feature. The results outperformed HOOF and LBP extracted from the apex frame.

Meanwhile, Xu et al. [13] use the same feature from their ME detection method for recognition purpose, i.e. Facial Dynamics Map (FDM) that estimates the movement between frames using optical flow estimation and classify movements of ME sequences. The authors increased computational efficiency by designing the FDM in parallel algorithm. It was also claimed that FDM will provide a meaningful representation by observing dynamic movement, compared to LBP variants.

Main Directional Mean Optical–flow (MDMO) was introduced by Liu et al. [96] for ME recognition. The MDMO is a local-based optical flow extracted from local region of interest. The alignment method is applied in optical flow to handle small head movements. MDMO is also insensitive to illumination and reports better performances compared to LBP-TOP and HOOF in CASME, CASME II, and SMIC dataset. On the other hand, Oh et al. [97] use multi-scale Riesz wavelet representation as feature for ME recognition, as it could capture monogenic signal components. Note that Riesz wavelet is inspired by phase-based method for computing optical flow.

Compared to LBP features, this category places more focus on the temporal dynamic variations. Several state-of-the-art methods in this category claim that illumination changes can be subjugated. However, head-pose variation could affect recognition rates. Therefore, face registration plays an important role when using any method of this category.

3.1.3 Gradient-based feature

There are several features that utilize gradient concepts to represent subtle motions. Gradients of an image, i.e. X, Y, and T derivatives, are used to compute the magnitude of gradient. This magnitude is usually large around the edges and corners which indicates huge intensity changes.

The first gradient-based descriptor implemented in ME recognition is the 3D-gradient descriptor [24]. In brief, 3D-gradients orientation histogram descriptor is generated from 12 region of interests (ROI) of the face. The gradient magnitudes and its respective orientations are obtained by its partial derivative along the X, Y, and T directions locally. Histograms representing shapes, vertical changes, and horizontal changes of every frame are constructed based on magnitudes and orientations excluding low magnitude vectors. The histograms are concatenated to a final descriptor for AU classification. The consideration of three phases of ME motion in this feature, which are constriction of the facial muscle, muscle construction, and muscle release, is useful in physiological analysis. However, it does not consider head-pose variation factor and that posed emotions were used for testing purpose.

In addition, Histograms of Oriented Gradients (HOG) [5] is one of the mainstream low-level features for ME recognition. According to [5], a 2D HOG can be formulated on the XY plane of an image. The first step is to obtain the horizontal and vertical derivatives through filtering of image with kernels. Next, the directions of gradient and its magnitudes are calculated to construct the histogram. If the gradient direction is in between two bins of the histogram, the vote by the pixel is split evenly into two bins. In short, based on the resulting gradient computation, a quantized orientation channel is constructed based on the weight of each pixel inside a block.

Another approach introduced in [5] is Histogram of Image Gradient Orientation (HIGO). Compared to HOG, the complexity of HIGO is lower and is a simplified version of HOG. Instead of weighted vote in HOG, it uses a simple vote to compute histogram bins result. By disregarding the first-order derivatives magnitude, the effect of illumination and contrast can be reduced. In addition, illumination does not affect HIGO which is statistically found in image cuboid. In the case of spatiotemporal feature extraction, the corresponding 3D HOG and 3D HIGO was extended from 2D HOG and 2D HIGO with three-orthogonal planes instead of *XY* plane used in 2D approaches. Note that the gradient intensity is usually normalized to counter variations in illumination or contrast. Experiment is then conducted using both approaches and both results outperform LBP-TOP in CASME II dataset. SMIC dataset contains a few types of data and both methods outperform LBP-TOP except when the samples used are infrared data. This is because of the textures from an infrared camera that are unlike the usual RGB videos with respect to illumination sensitivity.

Generally, gradient-based features depend heavily on gradient changes at pixel level. This category of features may have similar drawbacks to the LBP family of features, i.e. head-pose variation problems [5]. Nevertheless, some features are purposefully designed to combat illumination changes, i.e. HIGO [5]. This category of feature evolved from holistic feature to local feature extractions.

3.1.4 Generic features

There are several features that do not belong to any of the aforementioned categories. Nevertheless, there are a few techniques that are based on the concept of LBP-TOP. For example, Local Spatiotemporal Directional Features (LSTD) [47] extracts the binary code in a way similar to LBP-TOP. In [47], robust PCA (RPCA) and LSTD are implemented together to obtain subtle motion and local textures for recognition based on the local regions, with each region consisting of AU. The LSTD is extracted by obtaining the binary code from these regions via thresholding of the middle pixel with two neighbouring pixels. Similar to LBP-TOP, the binary codes are also obtained in three planes. The accuracy achieved by LSTD when conducted in CASME II dataset is about the same as the LBP-TOP approach when the radius in *X* and *Y* direction is small. However, as the radius increases, the LSTD method outperforms the LBP-TOP. Note that the performance evaluation in this method is more dependent on AU classification, instead of micro-emotions classification.

On the other hand, LBP-TOP also inspired the delivery of Spatiotemporal Completed Local Quantization Patterns (STCLQP). Huang et al. [51] introduced the use of STCLQP that considers other beneficial information that LBP-TOP disregards. To execute STCLQP, Completed Local Quantization Patterns (CLQP) is first implemented by extracting important components, for instances sign, magnitude, and orientation components. Each of the components in the spatial and temporal domains utilizes effective vector quantization method and produces a compact and discriminative codebook selection. The robust and compact codebook stores the binary pattern and assigns local pattern into its corresponding index. In addition, it is built offline by means of mapping unique local pattern to the nearest clustering centres for fast execution. The CLQP is further developed into STCLQP by implementation in three different planes. This technique does not need parameters tuning, but is much more complex compared to the conventional LBP-TOP.

Gabor filter is another well-known method to extract texture information from an object and is also used as a feature for ME recognition [98]. A Gabor wavelet displays attributes of spatial locality and orientation selectivity [99]. In addition, it could be contained in both space and frequency domains for local analysis. Slight changes in pose and any tiny alignment mistakes are tolerable in Gabor wavelet due to its robust characteristics. Hence, 2D Gabor filter and sparse representation (2DGSR) [98] are proposed for ME recognition. The convolution between Gabor kernel and facial image yields a Gabor representation. All Gabor representations are concatenated to each other to give an augmented feature vector which would accept face image local deformation to a limit and enhance face feature. Based on the results obtained in [98], the 2DGSR is compared with LBP-TOP and HOOF approaches where it is rated higher in terms of accuracy than other methods in both CASME and CASME II datasets. An advantage of the 2DGSR is its Gabor feature that utilizes information on spatial and orientation while maximizing the whole process' sturdiness. However, the recognition duration is subpar in comparison and the transformation speed requires improvement.

ME recognition with discriminant tensor subspace analysis (DTSA) performed by Wang et al. [74] deals with second-order tensor which consists of grey image. In order to decrease the dimensions required, two-sided transformations were implemented. Third-order DTSA is further extended for ME video data to produce discriminative features that improve classification results since it maintains spatial structure information. In addition, DTSA also maintains the local structure of samples distribution.

Another tensor-based feature was introduced by Ben et al. [73] that achieves maximum margin projection with tensor representation (MMPTR). In [73], ME is treated as third-order tensors where MMPTR is used to find transformation matrices for tensor data by Laplacian scatter. Through transformation or projection, the discriminative and geometry-preserving features are obtained for recognition. Comparatively to DTSA, the tensor size or MMPTR is larger and classification accuracy is 27.8% higher after tensor vectorization.

Additionally, Kamarol et al. [100] proposed spatiotemporal texture map (STTM) to represent micro-expression. The proposed work combines second-order moment and Harris Corner function in the extraction of spatial and temporal information to construct a histogram for SVM classification. The algorithm achieves superior results with selected samples and classes from CASME II. Further to this, Lu et al. [101] treated facial texture variations as muscle movement using Delaunay-based temporal coding model (DTCM). Instead of utilizing feature points as features, this technique extracts spatial information from Delaunay triangulation and temporal information from local subregions. Later, Sariyanidi et al. [102] introduced the first unsupervised learnt dynamic representation to analyse characteristics of facial expression variation based on localized movement where each coefficient represents a particular movement and the magnitude represents the intensity of the movement. The trained model is obtained from Gabor phase shift and achieves only 65.64% of recognition accuracy.

As a result, numerous numbers of low-level features have been proposed in different domains, regardless of spatial, dynamic, or spatial–temporal nature. Some environmental challenges are expected to be solved in earlier stages, i.e. face registration or pre-processing, while spontaneous subtle motions are expected to be handled by dynamic or temporal features with future improvements in performance accuracies. In order to provide more avenues for investigation and research beyond low-level representations, there is a need for public datasets which will facilitate higher-level representations with more samples.

3.2 Mid-level representation

As presented in Sect. 3.1, numerous efforts have been made to represent ME using low-level features. Most existing methods simply concatenate the spatial and temporal features of holistic or different regions of faces to perform classification. However, low-level features remain inadequate in representing subtle motions due to short duration, low intensity, noise and head-pose changes [103]. Therefore, mid-level features are introduced by turning low-level representations into richer features with stronger discriminative power [103].

Mid-level feature is a technique to transform local features into image representations for classification purpose where weightage is added to bring explicit meanings and knowledge to local features. Descriptor extracted in this category will be expressed as visual word content with image-level information, instead of a simple visual cue without explicit meaning. The most common mid-level technique is the bag-of-words (BoW) representation that is commonly used in affect recognition [104].

Nevertheless, mid-level representations used to describe ME are insufficient. To date, only two works were done to implement mid-level features in ME recognition. He et al. [103] proposed a multi-task mid-level feature learning (MMFL) mechanism to enhance discriminative factor. In this technique, each video clip containing ME is partitioned into several non-overlapping blocks and low-level features are extracted for local region representation. Class-specific feature mappings are performed using multi-task learning mechanism for each block. Local features with similar specific class are pulled and categorized together, while local features from different classes will be separated farther. In other words, local features are mapped into mid-level classes based on similarity, with a common mapping to link mutual information between classes. The projections of each class-specific mapping are calculated and concatenated to form a holistic feature representation. Subsequently, two weighting schemes calculated using mean optical strains are added to the concatenated feature, where one weight is used to highlight the active face region, while another focuses on inactive face region to avoid bias. The final holistic mid-level features within the same class are then fed into SVM for recognition.

The K-SVD algorithm proposed by Zheng et al. [105, 106] learns the sparse dictionary for mid-level feature. In this work, different sparse coefficients, such as LBP-TOP and HOOF are used to represent subtle face changes. A relaxed K-SVD algorithm (RK-SVD) is used for dictionary learning while minimizing the variance of sparse coefficients which were added to enhance similarity of the same class and the discriminative factor between other classes. The dictionary is updated iteratively until convergence is achieved where recognition is done using linear classification. As a conclusion, RK-SVD is used as mid-level classifier to recognize low-level features.

To conclude, there are limited mid-level representations proposed to handle ME recognition. Implementation of AU in solid mid-level representations, i.e. bag-of-words could be another future direction in this domain.

3.3 High-level representation

A high-level representation can be defined as a set of semantic data that are human interpretable, where the high-level features are a combination of several low-level features. For example, in facial authentication, a set of semantic information, i.e. age, gender, and skin colour, are used for authentication purpose, where all semantic data can be extracted from low-level features, such as colour or texture. The conventional low-level features are also called handcrafted features [2], extracted from pixels and fed into the classifier for identification. On the other hand, high-level approach leans towards feature learning, where the algorithm learns and understands useful features independently from raw image input. Due to emerging requirements in visual inspection, high-level representations have garnered significant attention in the computer vision community.

Recently, deep learning has attracted a huge attention due to its excellent performance in various domains, i.e. action recognition [107], scene text identification [108], medical image segmentation [109], etc. Not surprisingly, deep learning algorithms have been implemented in facial expressions and ME recognition. The state-of-the-art high-level representations in ME are extracted from convolutional neural network (CNN) in deep learning algorithm. Lower layers in convolutional neural network are used to extract low-level features, while higher layers extract high-level representation from huge number of labelled samples as input. Several works were done on the use of deep learning to extract high-level representations from low-level ones in micro-emotions recognition.

Notably, Patel et al. [79] are the first to apply CNN model in ME recognition. However, limited number of ME datasets restricts the feasibility of CNN, since it requires massive data for training. The authors used ImageNet-Vgg-f CNN pre-trained network [110] for transfer learning. Meanwhile, another CNN network was also trained based on CK+ and SPOS facial expression dataset. Assuming that normal facial expressions are similar to ME, features are selected and taken from the layer before the last fully connected layer. The authors also performed concurrent spatiotemporal feature computation and extended the work by adding evolutionary algorithms to search optimal deep features without overfitting. The proposed work is applied to CASME II dataset with higher accuracy than the LBP-TOP approach but is insufficient to match STCLQP. This could be due to data overfitting by training correctness caused by fitness function evaluation heuristic

To cope with the challenge for the lack of data in CNN training, Takalkar and Xu [81] extended the available datasets, i.e. CASME and CASME II through data augmentation to produce additional synthetic images from these datasets. Images in these datasets were flipped vertically to generate synthetic data and subsequently cropped and aligned as input to CNN. The deep network consists of convolution to extract features, rectified linear unit (ReLU) to rectify negative value, pooling to reduce dimension, and last but not least classification. Generally, convolution and pooling layers are defined as the feature extraction stage, while remaining fully connected layer acts as the classifier. The proposed work does not take temporal information into consideration and gives a slightly superior performance to the conventional LBP-TOP + SVM method. The authors further state that the imbalanced datasets could bring about bias effect to the results, and mislabelled annotation could confuse the network.

Additionally, Mayya et al. [89] introduced another high-level based representation by combining temporal interpolation with DCNN for recognition. In their work, all videos are interpolated to the same dimension prior to DCNN in CUDA. High-level features are then extracted using DCNN and sent to SVM for classifications. The ImageNet CNN in CAFFE is used for feature extraction in graphics processing unit (GPU) for faster performance. This was implemented in SMIC and CASME II datasets with only selected classes. Some classes, i.e. “Others” in CASME II with large amount of samples, are not used for training to prevent network confusion.

A Dual Temporal Scale Convolution Neural Network (DTSCNN) to recognize spontaneous facial micro actions was proposed by Peng et al. [111]. At its core, DTSCNN is a simplified two-stream network. In this work, different streams of DTSCNN are used to handle multiple frame rate of ME videos from dataset. An independent shallow network is present in every stream of DTSCNN to cope with overfitting. Furthermore, optical flow sequences are used as inputs to the network so that high-level features can be produced by a shallow network. In this work, spatial information and temporal information are considered. After the learning process, a linear SVM classifies output features. The work is tested in the merged dataset of CASME and CASME II. Results shown in this work indicate that the proposed method may outperform STCLQP, MDMO, and FDM with approximately 10% accuracy improvement.

A combination of CNN and long short-term memory (LSTM) was proposed by Kim et al. [80] to deal with spatial and temporal information. In spatial domain representative, each expression stage, i.e. onset, offset, or apex, is learned by the network instead of utilizing movement intensities. The variation in expressions classes, state, and state continuity is taken into account which makes the feature robust to illumination changes, while temporal characteristics of extracted spatial information from CNN are learned using LSTM. The LSTM neural network can be adopted to learn temporal information from video clips of different frame rates. The proposed techniques achieved remarkable results compared to the conventional LBP and its variants. However, the imbalanced samples in dataset continue to affect the confusion matrix results in [80].

Recently, Yu et al. [28] introduced a deeper cascaded peak-piloted network (DPCN) to recognize weak expressions. In this work, the authors used the peak expressions (apex) to supervise the non-peak expressions (onset/offset). Back propagation algorithm and cascaded fine-tune were added with improved performance and avert overfitting at the same time. However, the proposed network is tested in facial macro-expression dataset (i.e. CK+ dataset) instead of micro-expression dataset as highly expressive frames used to train the network must be strong enough to supervise weak expressions. Otherwise, the improvement is not significant. The best recognition accuracy achieved by this network in CK+ is 99.6%.

Overall, the features used in ME have evolved from low-level to high-level representations. However, limitations in high-level representations such as scarcity of data forcefully restrict the strength of deep learning approach. Different approaches have been proposed to overcome this limitation, such as transfer learning [79] or augmentation of available data [81]. However, the accuracy is still far from saturation. This presents a motivation and room for improvement in high-level approach for spontaneous ME recognition.

4 Micro-expression spotting

ME spotting is a stage where frames containing emotions are detected in time for a given video. There are less published works on ME spotting [112, 113] compared to ME recognition, and the approaches can be categorized into appearance-based, dynamic, and generic approach. The features used in ME spotting are often extracted using low-level representation as discussed in Sect. 3.1.

4.1 Appearance-based approach

Appearance-based approach normally refers to feature representation constructed in pixel-wise level, especially by the intensity value. Wu et al. [114] used the Gabor filter to spot ME in METT training data. METT training data are synthetic which makes spotting much easier compared to natural spontaneous reactions. Pfister et al. [46] used LBP-TOP as feature to detect ME and is tested in SVM, Multiple Kernel Learning (MKL), as well as Random Forest (RF) to achieve at least 70% detection rate. Moilanen et al. [115] spotted frames containing ME by computing feature difference (FD) using Chi-square distance of local binary pattern (LBP). Li et al. [5] also implemented ME spotting based on FD comparison. LBP and histogram of oriented optical flow (HOOF) are used as proof of concepts and that the LBP method outperforms HOOF. The authors are the first to propose a spotting algorithm from spontaneous long videos. However, this method is indifferent to non-micro-expression movements, such as eye blinking. Note that LBP-TOP and HOOF are further implemented for ME recognition as low-level representation, as discussed in Sects. 3.1.1 and 3.1.2, where HOOF is extended to FHOOF.

Besides intensity-based feature, methods such as 3D gradient histogram descriptor [29] or histogram of oriented gradients [116] are also reported for ME spotting. Davison et al. [117] used 3D-HOG to detect action units from FACS region locally. Contrary to [24, 29], Davison et al. focused only on areas of the face that contains specific AU. Features are extracted using 3D HOG in three-orthogonal planes to describe motion in each direction. Compared to LBP-TOP and HOOF, FACS-based 3D-HOG achieves better detection accuracy since this method ignores global emotions and emphasizes simple muscle movement, allowing complexity reduction. However, this method is inferior in terms of speed. Spatio–temporal Completed Local Quantization Patterns (STCLQP) was introduced by Huang et al. [51] to perform detection by extracting sign, magnitude, and orientation as feature. Classification is done by comparing extracted information to the codebook. At the same time, 3D-HOG and STCLQP are also presented as low-level representation for ME recognition as discussed in Sects. 3.1.3 and 3.1.4.

On the other hand, Lu et al. [118] use the differences of integral projection for ME spotting purpose. The feature is obtained by computing the difference between integral projections of neutral frame and the frame with expressive emotion. Note that feature difference is computed from the specific region of interest instead of a holistic face. This method achieves 92.89% accuracy, but is sensitive to motions and illumination changes.

4.2 Dynamic approach

In this domain, the feature is constructed based on non-rigid motion changes of subtle expression where motion changes are extracted for spotting purpose. Optical flow and optical strain were first proposed to detect subtle motion in [119]. This strategy [119] calculates the motion and strain derived from optical flow in ME spotting. Similarly, Shreve et al. [25] post-processed flow vectors before computing the strain to decrease the errors generated when approximating optical flow. This method [119] is tested on the UFS dataset with 100% detection rate. However, the samples from this dataset are posed and non-spontaneous. The method is also unusable if head translation is large and fast [25, 119]. Furthermore, optical strain feature, optical strain weight, and the combination of these are also reported to spot spontaneous emotion in [92]. Further details on ME recognition using optical flow and its variants can be found in Sect. 3.1.2.

The facial dynamic map (FDM) introduced by Xu et al. [13] characterizes movement of facial component for ME identification with optical flow estimation. Pixel level alignment is done based on optical flow, and the aligned field is divided into several cuboids. The directions of cuboids are used to characterize facial dynamics. Such method can be used to prevent head-pose variation from deteriorating accuracy rate. Xu et al. also implemented FDM for ME recognition purpose in Sect. 3.1.2. Other than FDM, a main directional maximal difference (MDMD) analysis [120] was introduced to spot ME based on maximal difference of magnitude in the main direction of optical flow features where it capitalizes the main direction of optical flow and obtains higher accuracy of spotting [120] in long videos. A deep learning method investigated by Li et al. [121] uses HOOF to detect the direction of facial muscle movements. The authors implemented a multi-task learning to localize facial landmarks and region-based normalized HOOF features. The algorithm was tested in CASME [6] dataset with detection rate of 80%. To date, this is the only approach that uses high-level technique for ME spotting.

A recent work published in 2018 [122] suggested the Riesz Pyramid for ME spotting. Riesz Pyramid was originally proposed to magnify subtle motions in video due to its good proxy for motion. Meanwhile, Duque et al. [122] use Riesz Pyramid to detect subtle motion by computing the amplitude and quaternionic phase from different subbands generated from input image using Laplacian Pyramid. The best level that represents subtle motion is chosen for detection out of different subbands. This approach is able to differentiate eye movement and ME. Riesz wavelet is also used for ME recognition, as shown in Sect. 3.1.2.

4.3 Generic approach

In this section, we distinguish features which are not variants of aforementioned approaches into generic categories. Recently, a micro-expression spotting benchmark (MESB) was introduced by Hong et al. [112] and Tran et al. [113]. The authors combined multi-scale sliding window to spot ME and provide a standardized performance evaluation of ME spotting method. In their work, it is assumed that spotting is a binary classification task, and the first type of measurement is to quantify direct performance of binary classification based on sliding windows (video containing ME or non-ME). The second measurement was made based on final outputs of detected windows. The authors considered false positive, true positive and missing samples in the second type of measurement and that the spotting method's saturation point is far from reach. Another spotting method, random walk model (RW) proposed by Xia et al. [53] is based on a probabilistic framework to detect spontaneous movement from a video sequence. The proposed method estimates the probability of frames having expression through geometric deformation. However, the accuracy is heavily depended on landmark localizations and multiple frames correlations

To our knowledge, most of the aforementioned methods are only tested in publicly available datasets, where the data are collected in controlled scenes. As mentioned, ME spotting in natural scenes with different environmental factors is still an open issue that is unresolved. Furthermore, all public datasets are presented in short duration video. As a result, the number of research works on detection is restricted to these datasets, in which, several techniques cannot be further tested in the context of long-duration video [17] with the exception to MDMD where the accuracy remains far from saturation [120]. A comprehensive ME dataset containing long video and natural scene is essential for further research in this area. We conclude all related ME spotting works in Table 2. Note that some approaches used classifiers to spot emotion and leave-one-subject-out (LOSO) for performance validation, while other methods omit classifier to detect emotion and only utilizes area under curve (AUC) for performance measure.

5 Recognition results and discussion

5.1 Result

Leave-one-video-out (LOVO) and leave-one-subject-out (LOSO) cross-validation are the most common methods used in ME recognition performance measurement. LOVO is operated by taking out one video sample for testing, while the remaining samples are used for training. The process is repeated for all samples in the dataset, and the mean accuracy is calculated from all computed accuracies. Hence, LOVO is sometimes called leave-one-sample-out (LOO). LOSO, on the other hand, reserves all sample videos belonging to one subject as testing data, while the remaining videos are used as training data [32]. The process is repeated in all subjects, and the mean accuracy is calculated as final result. Instead of LOVO and LOSO, *k*-fold validation [100, 111], repeated random sub-sampling validation [74], or basic hold-out methods are also used for performance evaluation with recall and precision graphs reported in other works [34, 97].

Most experiments are conducted on feature extraction and classification approaches with varying parameters and datasets. The results of previous works are extracted from relevant articles and compiled in Table 3 with regard to which feature, classifier, dataset and emotion classes were used to perform detection and recognition.

In this table, the highest achievement from respective papers is recorded and dedicated for low-level approaches due to its popularity in literature. A comparison of mid-level features and high-level features is compiled separately in Table 4 along with the performance protocols of each method.

From Tables 3 and 4, we observe that different approaches are tested in varying datasets. LBP-TOP is always used as a baseline in most studies; therefore, we include only results of LBP-TOP from [7, 8]. Some classes in dataset are purposely merged to overcome imbalance issues, such as 2DGSR [98], TICS + LBP [50], or FHOOF and FHOFO [72]. Among these approaches, STTM in low-level approach achieved superior results with 98.61% accuracy in CASME II. This is because

Table 2 Comparison of ME spotting approaches, accuracy, and dataset from different studies

Category	Approach	Classifier	Dataset	Accuracy (%)	Evaluation metric
Appearance-based approach	Gabor filter [114]	GentleSVM	METT	95.83	Tenfold cross-validation
	LBP-TOP [46]	SVM	York-DDT	65.0	LOSO
		MKL		67.0	
		MKL + TIM10		83.0	
		SVM + TIM10	SMIC (down-sampled to 25fps)	65.0	
		MKL + TIM10		70.3	
		RF + TIM20		78.9	
	Chi-square of LBP [115]	–	SMIC	71.0	ROC curve
		–	CASME	66.0	
	Feature difference of LBP [5]	–	SMIC	83.32	ROC curve
		–	CASME II	92.98	
	Feature difference of HOOF [5]	–	SMIC	69.41	ROC curve
		–	CASME II	64.99	
	3D gradient histogram (AU based) [29]	K-mean	Own dataset	95	–
	Histogram of oriented gradient [116]	–	In-house dataset	Recall—84.29 Precision—76.72	Precision and $F1$-measure
	3D-HOG [117]	–	SAMM	Recall 68.04	Recall, Precision, $F1$-measure, accuracy
		–	CASME II	F-measure—57.54	
	STCLQP [51]	SVM	SMIC	70.73	LOSO
	Integral projection [118]	–	CASME	84.80	Area Under Curve (AUC)
			CASME II	92.89	
Dynamic approach	Optical flow and strain [25, 119]	–	USF video collection	100	–
	Optical strain (OSF) [92]	SVM	SMIC	66.16	LOSO
	Optical strain weight (OSW) [92]	SVM	SMIC	63.72	LOSO
	OSF + OSW [92]	SVM	SMIC	72.87	LOSO
	FDM [13]	SVM (RBF)	SMIC (v1)	75.66	LOSO
		SVM (RBF)	SMIC	75.30	LOSO

Table 2 continued

Category	Approach	Classifier	Dataset	Accuracy (%)	Evaluation metric
	MDMD [120]	–	CAS(ME)^2	Recall—31.90 Precision—35.21 $F1$-measure—33.48	Recall, Precision, $F1$-measure, accuracy
	Deep learning with HOOF [121]	SVM	CASME	80	LOSO
	Riesz pyramid [122]	–	SMIC	89.8	AUC
			CASME II	95.13	
Generic approach	Multi-scale Sliding Window [112, 113]	SVM (LINEAR)	SMIC	66.33 (mean accuracy)	LOSO (false positive per window)
	Random walk [53]	Adaboost	SMIC	86.93	AUC
			CASME	92.08	

Table 3 Comparison of feature, classifier, and accuracy from different studies utilizing low-level features

Category	Feature	Classifier	Dataset	Number of classes	Accuracy (%)	Protocol
LBP family	LBP-TOP [7, 8]	SVM	SMIC	3	48.78	LOSO
			CASME II	5	63.41	LOSO
	LBP-SIP [32]	SVM (RBF)	SMIC	3	50/64.02	LOSO/LOVO
		SVM (poly/linear)	CASME II	5	44.53/66.40	LOSO/LOVO
	LBP-MOP [32]	SVM (RBF/linear)	SMIC	3	50.61/60.98	LOSO/LOVO
			CASME II	5	45.75/66.8	LOSO/LOVO
	CBP-TOP [76]	ELM	CASME	3	82.07	–
	LBP-TOP + Gaussian Jet [48]	SVM (poly)	SMIC (first version)	2 (pos/neg)	67.8	LOSO
	STLBP-IP [54]	SVM (Chi-square)	SMIC	3	57.93	LOSO
			CASME II	5	59.51	LOSO
	STLBP-IP [82]	FQRC	CASME II	5	79	LOSO
	STLBP-RIP [52]	KNN	SMIC	3	60.37	LOSO
		SVM (linear)	CASME	4	59.06	LOSO
		SVM (Chi-square)	CASME II	5	62.75	LOSO
	DISTLBP-RIP [52]	KNN	SMIC	3	63.41	LOSO
		SVM (linear)	CASME	4	64.33	LOSO
		SVM (Chi-square)	CASME II	5	64.78	LOSO
	TICS + LBP-TOP [50]	SVM (linear)	CASME	4 (merged)	61.86	LOSO

Table 3 continued

Category	Feature	Classifier	Dataset	Number of classes	Accuracy (%)	Protocol
			CASME II	5 (136 samples)	61.76	LOSO
	CIELuv + LBP-TOP [49]	SVM	CASME	4 (merged)	61.86	LOSO
			CASME II	4 (merged)	62.3	LOSO
	CIELab + LBP-TOP [49]	SVM	CASME	4 (merged)	61.86	LOSO
			CASME II	4 (merged)	61.11	LOSO
Optical flow family	Optical strain [91]	SVM (linear)	SMIC	3	53.56	LOSO
	OSF [34, 91]	SVM (linear)	SMIC	3	41.6	LOSO
			CASME II	5	51.01	LOVO
	OSW [34]	SVM (linear)	SMIC	3	49.39	LOSO
			CASME II	5	61.94	LOVO
	OSF + OSW [34]	SVM (linear)	SMIC	3	52.44	LOSO
			CASME II	5	63.16	LOVO
	FHOOF [72]	SVM (linear)	SMIC	3	44.51	LOSO/LOVO
			CASME	4 (merged)	63.45	LOSO/LOVO
			CASME II	4 (merged)	55.08	LOSO/LOVO
	FHOFO [72]	SVM (linear)	SMIC	3 (merged)	51.22/56.10	LOSO/LOVO
			CASME	4 (merged)	65.99/71.57	LOSO/LOVO
			CASME II	4 (merged)	55.86/64.06	LOSO/LOVO
	FDM [13]	SVM (RBF)	SMIC (v1)	2	71.43	LOSO
			SMIC	3	54.88	LOSO
			CASME	8	42.02	LOSO
			CASME II	7	41.96	LOSO
	MDMO [96]	SVM (poly)	SMIC(v1)	2	80	LOSO
			CASME	4	75.45	LOVO
			CASME II	4 (merged)	67.37	LOSO
	BI-WOOF [95]	SVM (linear)	SMIC	3	62.20	LOSO
			CASME II	5	58.85	
			CAS(ME)^2	3	–	
	Riesz wavelet [97]	SVM (linear)	CASME II	5	46.15	LOSO
Gradient-based approach	3D gradient [24]	*K*-mean	–	AU based	95	–
	HOG [5]	SVM (linear)	SMIC	3	57.93	LOSO
			CASME II	5	57.49	LOSO
	HIGO [5]	SVM (linear)	SMIC	3	65.24	LOSO
			CASME II	5	57.09	LOSO
Generic approach	LSTD [123]	–	SMIC	3	68.2927	LOSO
		–	CASME II	5	65.4472	LOSO
	DLSTD [123]	–	SMIC	3	68.2927	LOSO
		–	CASME II	5	63.4146	LOSO
	STCLQP [51]	SVM (linear)	SMIC	3	64.02	LOSO
			CASME	4	57.31	LOSO

Table 3 continued

Category	Feature	Classifier	Dataset	Number of classes	Accuracy (%)	Protocol
			CASME II	5	58.39	LOSO
	2DGSR [98]	Sparse representation classifier	CASME	4 (merged)	71.19	LOSO
			CASME II	4 (merged)	64.88	LOSO
	MMPTR [73]	–	CASME	4	ARR = 80.2	Hold-out
	DTSA [74]	ELM	CASME	5	46.9	Repeated random sub-sampling model
	STTM [100]	SVM (poly)	CASME II	5 (exclude "others")	98.61	Twofold validation
	DTCM [101]	SVM (linear)	SMIC(v1)	2	82.86	LOSO
	DTCM [101]	RF	CASME	4 (merged)	64.95	LOSO
		SVM (linear)	CASME II	4 (exclude "others")	72.06	LOSO
	Learnt model based on Gabor phase shift [102]	SVM (poly)	SMIC	AU based	65.64	LOSO

the largest class, i.e. "others" in CASME II, is not considered in their classification.

To date, the achievement of recognition accuracy is around 40% to a mere 83%, considering all classes in datasets involved in classification. For methods tested in all classes within the most challenging dataset of CASME II, the highest accuracy was only 79.89% in sparse coding method [105]. As mentioned, the uneven distribution samples among classes negatively impact recognition rate. Also, subtle and spontaneous movement of the face may limit accuracy. The trend of ME recognition is changing from low-level handcrafted feature to high-level approaches. However, the development of high-level approach is restricted by small dataset sizes.

Hence, augmentation of data or transfer learning is done to provide higher number of samples. Research work with testing on CAS(ME)^2 dataset and SAMM dataset is at the moment, lacking.

5.2 Discussion and future recommendation

In Sect. 1, we identify the challenges in ME recognition process. The first challenge is environmental factors, which are illumination changes and head-pose variation. To overcome head-pose variation works on face registration, faces in all frames can be aligned and normalized into the same position and size. Features such as FDM [13] are able to overcome head-pose changes. As far as illumination change is concerned, all algorithms were heavily tested in controlled and even illumination. Features like FHOOF [72], however, are not affected by illumination.

The second challenge is the subtle and spontaneous movement of facial component that deteriorates accuracy. Hence, Euler magnification [21] is introduced in the pre-processing stage to amplify low- intensity movements. Unfortunately, there is limited work done on magnification with accuracies reported at only 40–80%.

As for the third challenge, to date, there are several well-known datasets proposed on spontaneous response recognition. However, there are a few open issues that remain:

- All data are collected from a frontal view of faces. Is it possible to detect and recognize ME from side face profiles?
- All experiments are done in controlled environments, and the tested approaches may not be applicable in a natural environment with challenging factors, i.e. uneven illumination, noise, etc. ME datasets focusing on real-world environment are required.
- Uneven distribution and small sample sizes deteriorate the accuracy of tested algorithms. A dataset with large size of even samples among classes is essential for high-level approach.
- All datasets are taken based on facial expressions only. Body language MEs may be helpful in enhancing accuracy, since it is proven that similarity of body language can be cross-cultured [124].

Table 4 Comparison of feature, classifier, and accuracy from various mid-level and high-level feature studies

Category	Feature/method	Classifier	Dataset	Number of classes	Accuracy (%)	Protocol
Mid-level representation	MMFL [103]	SVM (RBF)	SMIC	3	63.15 (LBP-MOP based)	LOSO
			CASME II	5	59.81 (LBP-MOP based)	LOSO
	Sparse coefficient coding [105]	RK-SVD	CASME	4 (merged)	69.04 (LBP-TOP)	LOSO
			CASME II	5	79.89 (LBP-TOP based)	LOSO
High-level representation	CNN (selective) [79]	SVM (linear)	SMIC	3	53.6	LOSO
			CASME II	5	47.5	LOSO
	CNN (augmenting data) [81]	Fully connected layer	CASME	6 (extend)	74.25	80% training, 10% testing, 10% validation
			CASME II	6 (extend)	75.57	
			CASME + CASME II (merged)	6 (extend)	78.02	
	DCNN + CUDA [89]	Linear SVC	SMIC	3	65.85/78.05	LOVO/75% training, 25% testing
			CASME II	5	64.9/67.742	
	DTSCNN [111]	SVM (linear)	CASME + CASME II (merged)	4	66.67	Threefold cross-validation
	CNN + LSTM (spatial feature) [80]	Fully connected layer	CASME II	5	58.54	LOSO
	CNN + LSTM (spatial–temporal feature) [80]				60.98	
	DPCN [28]	Fully connected layer	CK + (weak expression in macro-expression)	6	99.6	Tenfold cross-validation

Also, based on this review we suggest the following ideas for future work in ME recognition.

- In pre-processing, works on magnification are lacking compared to feature extraction and classification. Enhancement of the intensities of subtle motions can be proposed for future work.
- Studies in ME spotting are limited. Mid-level or high-level approaches can be proposed to detect ME in long and real-world videos, instead of controlled environment.
- Most of the proposed features are focused on low-level approach with only two existing works on mid-level features. Mid-level representations such as bag-of-words can be proposed in ME recognition.

- High-level approaches, i.e. deep learning, require a lot of data for training purpose. Hence, comprehensive datasets are necessary to realize the potential of high-level approaches in ME recognition in the development of new datasets.

6 Conclusion

In this paper, we identify the challenges, trends, cutting edge methods, and provide recommendations for ME recognition. Firstly, common challenges in ME recognition processes are determined and associated with relevant solutions. A brief review of ME spotting was also included albeit less popular than feature extraction. In feature extraction, many low-level approaches have been proposed, but with unsatisfactory results. Feature representations are evolving from low-level approach to mid-level and high-level approach. The characteristics and limitations of publicly available ME datasets are highlighted and common performance protocols used for evaluation are identified. A better view of performance comparison along with possible directions to overcome drawbacks was also provided.

Acknowledgements This work is supported by Ministry of Higher Education (MOHE), Malaysia, under Fundamental Research Grant Scheme (FRGS) (Reference code: FRGS/1/2016/TK04/TARUC/02/1).

Compliance with ethical standards

Conflict of interest The author declares that they have no conflict of interest.

References

1. Zhang, M., Fu, Q., Chen, Y.-H., Fu, X.: Emotional context influences micro-expression recognition. PLoS ONE **9**(4), e95018 (2014)
2. Takalkar, M., Xu, M., Wu, Q., Chaczko, Z.: A survey: facial micro-expression recognition. Multimed. Tools Appl. **77**, 301 (2017)
3. Ekman, P.: Darwin, deception, and facial expression. Ann. N. Y. Acad. Sci. **1000**(1), 205–221 (2003)
4. Ekman, P.: Micro expression training tool (METT) and subtle expression training tool (SETT). Paul Ekman Company, San Francisco (2003)
5. Li, X., Xiaopeng, H., Moilanen, A., Huang, X., Pfister, T., Zhao, G., Pietikainen, M.: Towards reading hidden emotions: a comparative study of spontaneous micro-expression spotting and recognition methods. IEEE Trans. Affect Comput. **PP**(99), 1–1 (2017)
6. Wen-Jing, Y., Wu, Q., Yong-Jin, L., Su-Jing, W., Fu, X.: CASME database: a dataset of spontaneous micro-expressions collected from neutralized faces. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 22–26 April 2013, pp. 1–7
7. Yan, W.-J., Li, X., Wang, S.-J., Zhao, G., Liu, Y.-J., Chen, Y.-H., Fu, X.: CASME II: an improved spontaneous micro-expression database and the baseline evaluation. PLoS ONE **9**(1), e86041 (2014)
8. Li, X., Pfister, T., Huang, X., Zhao, G., Pietikäinen, M.: A spontaneous micro-expression database: inducement, collection and baseline. In: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 22–26 April 2013, pp. 1–6
9. Ekman, P., Friesen, W.V.: Facial Action Coding System, vol. 1. Consulting Psychologists Press, Washington (1978)
10. Tian, Y.I., Kanade, T., Cohn, J.F.: Recognizing action units for facial expression analysis. IEEE Trans. Pattern Anal. Mach. Intell. **23**(2), 97–115 (2001)
11. Frank, M., Herbasz, M., Sinuk, K., Keller, A., Nolan, C.: 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, New York City (2009)
12. Yan, W.-J., Wu, Q., Liang, J., Chen, Y.-H., Fu, X.: How fast are the leaked facial expressions: the duration of micro-expressions. J. Nonverbal Behav. **37**(4), 217–230 (2013)
13. Xu, F., Zhang, J., Wang, J.Z.: Microexpression identification and categorization using a facial dynamics map. IEEE Trans. Affect Comput. **8**(2), 254–267 (2017)
14. Ojala, T., Pietikäinen, M., Harwood, D.: A comparative study of texture measures with classification based on featured distributions. Pattern Recognit. **29**(1), 51–59 (1996)
15. Black, M.J., Anandan, P.: The robust estimation of multiple motions: parametric and piecewise-smooth flow fields. Comput. Vis. Image Underst. **63**(1), 75–104 (1996)
16. Shreve, M., Brizzi, J., Fefilatyev, S., Luguev, T., Goldgof, D., Sarkar, S.: Automatic expression spotting in videos. Image Vis. Comput. **32**(8), 476–486 (2014)
17. Qu, F., Wang, S.J., Yan, W.J., Li, H., Wu, S., Fu, X.: CAS(ME)^2: a database for spontaneous macro-expression and micro-expression spotting and recognition. IEEE Trans. Affect Comput. **PP**(99), 1–1 (2017). https://doi.org/10.1109/taffc.2017.2654440
18. Yan, W.-J., Wang, S.-J., Chen, Y.-H., Zhao, G., Fu, X.: Quantifying micro-expressions with constraint local model and local binary pattern. In: Cham 2015. Computer Vision—ECCV 2014 Workshops, pp. 296–305. Springer
19. Porter, S., Brinke, L.T.: Reading between the lies: identifying concealed and falsified emotions in universal facial expressions. Psychol. Sci. **19**(5), 508–514 (2008). https://doi.org/10.1111/j.1467-9280.2008.02116.x
20. Yan, W.-J., Wang, S.-J., Liu, Y.-J., Wu, Q., Fu, X.: For micro-expression recognition: database and suggestions. Neurocomputing **136**, 82–87 (2014)
21. Wang, Y., See, J., Oh, Y.-H., Phan, R.C.-W., Rahulamathavan, Y., Ling, H.-C., Tan, S.-W., Li, X.: Effective recognition of facial micro-expressions with video motion magnification. Multimed. Tools Appl. **76**(20), 21665–21690 (2017)
22. Ngo, A.C.L., Oh, Y.H., Phan, R.C.W., See, J.: Eulerian emotion magnification for subtle expression recognition. In: IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 20–25 March 2016, pp. 1243–1247
23. Le Ngo, A.C., Phan, R.C.-W., See, J.: Spontaneous subtle expression recognition: imbalanced databases and solutions. In: Cham 2015. Computer Vision—ACCV 2014, pp. 33–48. Springer
24. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor. In: 3rd International Conference on Imaging for Crime Detection and Prevention (ICDP 2009), 3–3 Dec 2009, pp. 1–6

25. Shreve, M., Godavarthy, S., Goldgof, D., Sarkar, S.: Macro- and micro-expression spotting in long videos using spatio-temporal strain. Face Gesture **2011**(21–25), 51–56 (2011)
26. Warren, G., Schertler, E., Bull, P.: Detecting deception from emotional and unemotional cues. J. Nonverbal Behav. **33**(1), 59–69 (2009)
27. Davison, A.K., Lansley, C., Costen, N., Tan, K., Yap, M.H.: SAMM: a spontaneous micro-facial movement dataset. IEEE Trans. Affect Comput. **PP**(99), 1–1 (2016)
28. Yu, Z., Liu, Q., Liu, G.: Deeper cascaded peak-piloted network for weak expression recognition. Vis. Comput. 1–9 (2017)
29. Polikovsky, S., Kameda, Y., Ohta, Y.: Facial micro-expression detection in hi-speed video based on facial action coding system (FACS). IEICE Trans. Inf. Syst. **E96.D**(1), 81–92 (2013)
30. Sariyanidi, E., Gunes, H., Cavallaro, A.: Automatic analysis of facial affect: a survey of registration, representation, and recognition. IEEE Trans. Pattern Anal. Mach. Intell. **37**(6), 1113–1133 (2015)
31. Sariyanidi, E., Gunes, H., Cavallaro, A.: Robust registration of dynamic facial sequences. IEEE Trans. Image Process. **26**(4), 1708–1722 (2017)
32. Wang, Y., See, J., Phan, R.C.W., Oh, Y.-H.: Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. PLoS ONE **10**(5), e0124674 (2015)
33. Le Ngo, A.C., See, J., Phan, R.C.W.: Sparsity in dynamics of spontaneous subtle emotions: analysis and application. IEEE Trans. Affect Comput. **8**(3), 396–411 (2017)
34. Liong, S.-T., See, J., Phan, R.C.-W., Le Ngo, A.C., Oh, Y.-H., Wong, K.: Subtle expression recognition using optical strain weighted features. In: Cham 2015. Computer Vision—ACCV 2014 Workshops, pp. 644–657. Springer
35. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001 2001, vol. 511, pp. I-511–I-518
36. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, 16–21 June 2012, pp. 2879–2886
37. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Cham 2014. Computer Vision—ECCV 2014, pp. 109–122. Springer
38. Matsugu, M., Mori, K., Mitari, Y., Kaneda, Y.: Subject independent facial expression recognition with robust face detection using a convolutional neural network. Neural Netw. **16**(5), 555–559 (2003)
39. Liao, S., Jain, A.K., Li, S.Z.: A fast and accurate unconstrained face detector. IEEE Trans. Pattern Anal. Mach. Intell. **38**(2), 211–223 (2016)
40. Ranjan, R., Patel, V.M., Chellappa, R.: A deep pyramid deformable part model for face detection. In: IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), 8–11 Sept 2015, pp. 1–8
41. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: a deep learning approach. In: IEEE International Conference on Computer Vision (ICCV), 7–13 Dec 2015, pp. 3676–3684
42. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7–12 June 2015, pp. 5325–5334 (2015)
43. Ranjan, R., Patel, V.M., Chellappa, R.: HyperFace: A deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1–1 (2017)
44. Yan, J., Zhang, X., Lei, Z., Li, S.Z.: Face detection by structural models. Image Vis. Comput. **32**(10), 790–799 (2014)
45. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models-their training and application. Comput. Vis. Image Underst. **61**(1), 38–59 (1995)
46. Pfister, T., Li, X., Zhao, G., Pietikäinen, M.: Recognising spontaneous facial micro-expressions. In: IEEE International Conference on Computer Vision (ICCV) 2011, pp. 1449–1456. IEEE
47. Wang, S.-J., Yan, W.-J., Zhao, G., Fu, X., Zhou, C.-G.: Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In: Agapito, L., Bronstein, M.M., Rother, C. (eds.) Computer Vision—ECCV 2014 Workshops: Zurich, Switzerland, September 6–7 and 12, 2014, Proceedings, Part I. pp. 325–338. Springer, Cham (2015)
48. Ruiz-Hernandez, J.A., Pietikäinen, M.: Encoding local binary patterns using the re-parametrization of the second order Gaussian jet. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 22–26 April 2013, pp. 1–6
49. Wang, S.J., Yan, W.J., Li, X., Zhao, G., Zhou, C.G., Fu, X., Yang, M., Tao, J.: Micro-expression recognition using color spaces. IEEE Trans. Image Process. **24**(12), 6034–6047 (2015)
50. Wang, S.J., Yan, W.J., Li, X., Zhao, G., Fu, X.: Micro-expression recognition using dynamic textures on tensor independent color space. In: 22nd International Conference on Pattern Recognition, 24–28 Aug 2014, pp. 4678–4683
51. Huang, X., Zhao, G., Hong, X., Zheng, W., Pietikäinen, M.: Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns. Neurocomputing **175**, 564–578 (2016)
52. Huang, X.H., Wang, S.J., Liu, X., Zhao, G., Feng, X., Pietikainen, M.: Discriminative spatiotemporal local binary pattern with revisited integral projection for spontaneous facial micro-expression recognition. IEEE Trans. Affect Comput. **PP**(99), 1–1 (2017)
53. Xia, Z., Feng, X., Peng, J., Peng, X., Zhao, G.: Spontaneous micro-expression spotting via geometric deformation modeling. Comput. Vis. Image Underst. **147**, 87–94 (2016)
54. Huang, X., Wang, S.J., Zhao, G., Piteikäinen, M.: Facial micro-expression recognition using spatiotemporal local binary pattern with integral projection. In: IEEE International Conference on Computer Vision Workshop (ICCVW), 7–13 Dec 2015, pp. 1–9
55. Liu, Q., Deng, J., Tao, D.: Dual sparse constrained cascade regression for robust face alignment. IEEE Trans. Image Process. **25**(2), 700–712 (2016)
56. Tzimiropoulos, G.: Project-out cascaded regression with an application to face alignment. In: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 7–12 June 2015, pp. 3659–3667
57. Xiong, X., Torre, F.D.L.: Supervised descent method and its applications to face alignment. In: 2013 IEEE Conference on Computer Vision and Pattern Recognition, 23–28 June 2013, pp. 532–539
58. Gines Hidalgo, Z.C., Tomas, S., Shih-En, W., Hanbyul, J., Yaser, S.: OpenPose Library. https://github.com/CMU-Perceptual-Computing-Lab/openpose (2017). 2018
59. Simon, T., Joo, H., Matthews, I., Sheikh, Y.: Hand keypoint detection in single images using multiview bootstrapping. In: CVPR 2017, p. 4
60. Deng, W., Fang, Y., Xu, Z., Hu, J.: Facial landmark localization by enhanced convolutional neural network. Neurocomputing **273**, 222–229 (2018)
61. Bian, P., Xie, Z., Jin, Y.: Multi-task feature learning-based improved supervised descent method for facial landmark detection. SIViP **12**(1), 17–24 (2018)
62. Pan, W., Qin, K., Chen, Y.: An adaptable-multilayer fractional fourier transform approach for image registration. IEEE Trans. Pattern Anal. Mach. Intell. **31**(3), 400–414 (2009)

63. Kumar, S., Azartash, H., Biswas, M., Nguyen, T.: Real-time affine global motion estimation using phase correlation and its application for digital image stabilization. IEEE Trans. Image Process. **20**(12), 3406–3418 (2011)
64. Xiong, X., Qin, K.: Linearly estimating all parameters of affine motion using radon transform. IEEE Trans. Image Process. **23**(10), 4311–4321 (2014)
65. Lucas, B.D.: Generalized image matching by the method of differences. (1985)
66. Baker, S., Matthews, I.: Lucas–Kanade 20 years on: a unifying framework. Int. J. Comput. Vis. **56**(3), 221–255 (2004)
67. Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: Robust and efficient parametric face alignment. In: International Conference on Computer Vision, 6–13 Nov 2011, pp. 1847–1854
68. Evangelidis, G.D., Psarakis, E.Z.: Parametric image alignment using enhanced correlation coefficient maximization. IEEE Trans. Pattern Anal. Mach. Intell. **30**(10), 1858–1865 (2008)
69. Pfister, T., Li, X., Zhao, G., Pietikäinen, M.: Differentiating spontaneous from posed facial expressions within a generic facial expression recognition framework. In: IEEE International Conference on Computer Vision Workshops (ICCV Workshops), 6–13 Nov 2011, pp. 868–875
70. Gunn, S.R.: Support vector machines for classfication and regression. ISIS Tech. Rep. **14**(1), 5–16 (1998)
71. Guo, Y., Tian, Y., Gao, X., Zhang, X.: Micro-expression recognition based on local binary patterns from three orthogonal planes and nearest neighbor method. In: 2014 International Joint Conference on Neural Networks (IJCNN), 6–11 July 2014, pp. 3473–3479
72. Happy, S.L., Routray, A.: Fuzzy histogram of optical flow orientations for micro-expression recognition. IEEE Trans. Affect Comput. **PP**(99), 1–1 (2017)
73. Ben, X., Zhang, P., Yan, R., Yang, M., Ge, G.: Gait recognition and micro-expression recognition based on maximum margin projection with tensor representation. Neural Comput. Appl. **27**(8), 2629–2646 (2016)
74. Wang, S.-J., Chen, H.-L., Yan, W.-J., Chen, Y.-H., Fu, X.: Face recognition and micro-expression recognition based on discriminant tensor subspace analysis plus extreme learning machine. Neural Process. Lett. **39**(1), 25–43 (2014)
75. Adegun, I.P., Vadapalli, H.B.: Automatic recognition of micro-expressions using local binary patterns on three orthogonal planes and extreme learning machine. In: Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech), 30 Nov 2016–2 Dec 2016, pp. 1–5
76. Guo, Y., Xue, C., Wang, Y., Yu, M.: Micro-expression recognition based on CBP-TOP feature with ELM. Opt. Int. J. Light Electron Opt. **126**(23), 4446–4451 (2015)
77. Izenman, A.J.: Linear discriminant analysis. In: Casella, G., Fienberg, S., Olkin, I. (eds.) Modern Multivariate Statistical Techniques, pp. 237–280. Springer, Cham (2013)
78. Fan, R.-E., Chang, K.-W., Hsieh, C.-J., Wang, X.-R., Lin, C.-J.: LIBLINEAR: a library for large linear classification. J. Mach. Learn. Res. **9**, 1871–1874 (2008)
79. Patel, D., Hong, X., Zhao, G.: Selective deep features for micro-expression recognition. In: 2016 23rd International Conference on Pattern Recognition (ICPR), 4–8 Dec 2016, pp. 2258–2263
80. Kim, D.H., Baddar, W., Jang, J., Ro, Y.M.: Multi-objective based spatio-temporal feature representation learning robust to expression intensity variations for facial expression recognition. IEEE Trans. Affect Comput. **66**(99), 1–1 (2017)
81. Takalkar, M.A., Xu, M.: Image based facial micro-expression recognition using deep learning on small datasets. In: International Conference on Digital Image Computing: Techniques and Applications (DICTA), 29 Nov 2017–1 Dec 2017, pp. 1–7
82. Lim, C.H., Goh, K.M.: Fuzzy qualitative approach for micro-expression recognition. In: Asic-Pasific Signal and Information Processing Association Annual Submit and Conference Kuala Lumpur, 12nd–15th Dec 2017
83. Lim, C.H., Risnumawan, A., Chan, C.S.: A scene image is nonmutually exclusive—a fuzzy qualitative scene understanding. IEEE Trans. Fuzzy Syst. **22**(6), 1541–1556 (2014)
84. Liu, H., Coghill, G.M., Barnes, D.P.: Fuzzy qualitative trigonometry. Int. J. Approx. Reason. **51**(1), 71–88 (2009)
85. Shen, Q., Leitch, R.: Fuzzy qualitative simulation. IEEE Trans. Syst. Man Cybern. **23**(4), 1038–1061 (1993)
86. Zhao, G., Pietikainen, M.: Dynamic texture recognition using local binary patterns with an application to facial expressions. IEEE Trans. Pattern Anal. Mach. Intell. **29**(6), 915–928 (2007)
87. House, C., Meyer, R.: Preprocessing and descriptor features for facial micro-expression recognition (2015)
88. Wang, Y., See, J., Phan, R.C.-W., Oh, Y.-H.: LBP with six intersection points: reducing redundant information in LBP-TOP for micro-expression recognition. In: Cham 2015. Computer Vision—ACCV 2014, pp. 525–537. Springer
89. Mayya, V., Pai, R.M., Pai, M.M.M.: Combining temporal interpolation and DCNN for faster recognition of micro-expressions in video sequences. In: International Conference on Advances in Computing, Communications and Informatics (ICACCI), 21–24 Sept 2016, pp. 699–703
90. Lee, S.H., Kim, H., Ro, Y.M., Plataniotis, K.N.: Using color texture sparsity for facial expression recognition. In: 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), 22–26 April 2013, pp. 1–6
91. Liong, S.-T., Phan, R.C.-W., See, J., Oh, Y.-H., Wong, K.: Optical strain based recognition of subtle emotions. In: International Symposium on Intelligent Signal Processing and Communication Systems (ISPACS) 2014, pp. 180–184. IEEE
92. Liong, S.-T., See, J., Phan, R.C.W., Oh, Y.-H., Cat Le Ngo, A., Wong, K., Tan, S.-W.: Spontaneous subtle expression detection and recognition based on facial strain. Signal Process. Image Commun. **47**, 170–182 (2016)
93. Ben, X., Jia, X., Yan, R., Zhang, X., Meng, W.: Learning effective binary descriptors for micro-expression recognition transferred by macro-information. Pattern Recogn. Lett. **107**, 50–58 (2017)
94. Chaudhry, R., Ravichandran, A., Hager, G., Vidal, R.: Histograms of oriented optical flow and Binet-Cauchy kernels on nonlinear dynamical systems for the recognition of human actions. In: IEEE Conference on Computer Vision and Pattern Recognition, 20–25 June 2009, pp. 1932–1939
95. Liong, S.-T., See, J., Wong, K., Phan, R.C.-W.: Less is more: micro-expression recognition from video using apex frame. Signal Process. Image Commun. **62**, 82–92 (2018)
96. Liu, Y.J., Zhang, J.K., Yan, W.J., Wang, S.J., Zhao, G., Fu, X.: A main directional mean optical flow feature for spontaneous micro-expression recognition. IEEE Trans. Affect Comput. **7**(4), 299–310 (2016)
97. Oh, Y.H., Ngo, A.C.L., See, J., Liong, S.T., Phan, R.C.W., Ling, H.C.: Monogenic Riesz wavelet representation for micro-expression recognition. In: IEEE International Conference on Digital Signal Processing (DSP), 21–24 July 2015, pp. 1237–1241
98. Zheng, H.: Micro-expression recognition based on 2D Gabor filter and sparse representation. J. Phys. Conf. Ser. **787**(1), 012013 (2017)
99. Chengjun, L., Wechsler, H.: Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition. IEEE Trans. Image Process. **11**(4), 467–476 (2002)
100. Kamarol, S.K.A., Jaward, M.H., Parkkinen, J., Parthiban, R.: Spatiotemporal feature extraction for facial expression recognition. IET Image Process. **10**, 534 (2016)

101. Lu, Z., Luo, Z., Zheng, H., Chen, J., Li, W.: A Delaunay-based temporal coding model for micro-expression recognition. In: Cham 2015. Computer Vision—ACCV 2014 Workshops, pp. 698–711. Springer
102. Sariyanidi, E., Gunes, H., Cavallaro, A.: Learning bases of activity for facial expression recognition. IEEE Trans. Image Process. **26**(4), 1965–1978 (2017)
103. He, J., Hu, J.-F., Lu, X., Zheng, W.-S.: Multi-task mid-level feature learning for micro-expression recognition. Pattern Recognit. **66**, 44–52 (2017)
104. Sikka, K., Wu, T., Susskind, J., Bartlett, M.: Exploring bag of words architectures in the facial expression domain. In: Berlin, Heidelberg 2012. Computer Vision—ECCV 2012. Workshops and Demonstrations, pp. 250–259. Springer, Berlin
105. Zheng, H., Geng, X., Yang, Z.: A Relaxed K-SVD algorithm for spontaneous micro-expression recognition. In: Cham 2016. PRICAI 2016: Trends in Artificial Intelligence, pp. 692–699. Springer
106. Zheng, H., Zhu, J., Yang, Z., Jin, Z.: Effective micro-expression recognition using relaxed K-SVD algorithm. Int. J. Mach. Learn. Cybern. **8**(6), 2043–2049 (2017)
107. Pei, L., Ye, M., Zhao, X., Dou, Y., Bao, J.: Action recognition by learning temporal slowness invariant features. Vis. Comput. **32**(11), 1395–1404 (2016)
108. Wu, H., Zou, B., Zhao, Y.-Q., Guo, J.: Scene text detection using adaptive color reduction, adjacent character model and hybrid verification strategy. Vis. Comput. **33**(1), 113–126 (2017)
109. Bi, L., Kim, J., Kumar, A., Fulham, M., Feng, D.: Stacked fully convolutional networks with multi-channel learning: application to medical image segmentation. Vis. Comput. **33**(6), 1061–1071 (2017)
110. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the Devil in the Details: Delving Deep into Convolutional Nets. CoRR abs/1405.3531 (2014)
111. Peng, M., Wang, C., Chen, T., Liu, G., Fu, X.: Dual temporal scale convolutional neural network for micro-expression recognition. Front. Psychol. **8**, 1745 (2017)
112. Hong, X., Tran, T.-K., Zhao, G.: Micro-Expression Spotting: A Benchmark. CoRR abs/1710.02820 (2017)
113. Tran, T.-K., Hong, X., Zhao, G.: Sliding window based micro-expression spotting: a benchmark. In: Cham 2017. Advanced Concepts for Intelligent Vision Systems, pp. 542–553. Springer
114. Wu, Q., Shen, X., Fu, X.: The machine knows what you are hiding: an automatic micro-expression recognition system. In: Berlin, Heidelberg 2011. Affective Computing and Intelligent Interaction, pp. 152–162. Springer, Berlin
115. Moilanen, A., Zhao, G., Pietikäinen, M.: Spotting rapid facial movements from videos using appearance-based feature difference analysis. In: 22nd International Conference on Pattern Recognition, 24–28 Aug 2014, pp. 1722–1727
116. Davison, A.K., Yap, M.H., Lansley, C.: Micro-facial movement detection using individualised baselines and histogram-based descriptors. In: IEEE International Conference on Systems, Man, and Cybernetics, 9–12 Oct 2015, pp. 1864–1869
117. Davison, A.K., Lansley, C., Ng, C.-C., Tan, K., Yap, M.H.: Objective micro-facial movement detection using FACS-based regions and baseline evaluation. CoRR abs/1612.05038 (2016)
118. Lu, H., Kpalma, K., Ronsin, J.: Micro-expression detection using integral projections. J WSCG **25**(2), 87–96 (2017)
119. Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D., Sarkar, S.: Towards macro- and micro-expression spotting in video using strain patterns. In: Workshop on Applications of Computer Vision (WACV), 7–8 Dec 2009, pp. 1–6
120. Wang, S.-J., Wu, S., Qian, X., Li, J., Fu, X.: A main directional maximal difference analysis for spotting facial movements from long-term videos. Neurocomputing **230**, 382–389 (2017)
121. Li, X., Yu, J., Zhan, S.: Spontaneous facial micro-expression detection based on deep learning. In: IEEE 13th International Conference on Signal Processing (ICSP), 6–10 Nov 2016, pp. 1130–1134
122. Duque, C., Alata, O., Emonet, R., Legrand, A.-C., Konik, H.: Micro-expression spotting using the Riesz pyramid. In: WACV 2018 (2018)
123. Wang, S.-J., Yan, W.-J., Zhao, G., Fu, X., Zhou, C.-G.: Micro-expression recognition using robust principal component analysis and local spatiotemporal directional features. In: Cham 2015. Computer Vision—ECCV 2014 Workshops, pp. 325–338. Springer
124. Parkinson, C., Walker, T.T., Memmi, S., Wheatley, T.: Emotions are understood from biological motion across remote cultures. Emotion **17**(3), 459–477 (2017)

Kam Meng Goh received the bachelor degree in 2010 and the Ph.D. degree in 2015 from University Teknologi Malaysia. He is currently a senior lecturer with Tunku Abdul Rahman University College in Kuala Lumpur, Malaysia. His current research interests include computer vision and recognition, artificial intelligence, and emotional information processing.

Chee How Ng graduated with Bachelor in Engineering (Honour) Electrical and Electronic from Tunku Abdul Rahman University College (TAR UC) in 2016. He is currently a research assistant and part-time tutor in TAR UC. His current research interests include computer vision, pattern recognition, and emotional information processing.

Li Li Lim graduated with an Advanced Diploma in Technology (Electronic Engineering) from Tunku Abdul Rahman College in 2004 and was subsequently awarded a MSc. in Electronics from Queen's University Belfast in 2005. She is now a full-time Senior Lecturer at the Faculty of Engineering at Tunku Abdul Rahman University College (TAR UC) and a member of the Centre for Multimodal Signal Processing at TAR UC. Her current research interest includes channel coding, MIMO symbol detection, and statistical signal processing for a wide application area.

U. U. Sheikh received his Ph.D. degree (2009) in image processing and computer vision from Universiti Teknologi Malaysia (UTM). He is currently active in research on computer vision and embedded systems design. He is currently a senior lecturer at Universiti Teknologi Malaysia, Skudai, Johor, Malaysia.