A Review of Automated Micro-expression Analysis

Koo Sie Min, Mohd Asyraf Zulkifley * & Nor Azwan Mohamed Kamari

Faculty of Engineering & Built Environment, Universiti Kebangsaan Malaysia, 43600 UKM Bangi, Selangor, Malaysia

**Corresponding author: asyraf.zulkifley@ukm.edu.my*

Received 24 September 2021, Received in revised form 23 February 2022
Accepted 22 March 2022, Available online 30 September 2022

ABSTRACT

Micro-expression is a type of facial expression that is manifested for a very short duration. It is difficult to recognize the expression manually because it involves very subtle facial movements. Such expressions often occur unconsciously, and therefore are defined as a basis to help identify the real human emotions. Hence, an automated approach to micro-expression recognition has become a popular research topic of interest recently. Historically, the early researches on automated micro-expression have utilized traditional machine learning methods, while the more recent development has focused on the deep learning approach. Compared to traditional machine learning, which relies on manual feature processing and requires the use of formulated rules, deep learning networks produce more accurate micro-expression recognition performances through an end-to-end methodology, whereby the features of interest were extracted optimally through the training process, utilizing a large set of data. This paper reviews the developments and trends in micro-expression recognition from the earlier studies (hand-crafted approach) to the present studies (deep learning approach). Some of the important topics that will be covered include the detection of micro-expression from short videos, apex frame spotting, micro-expression recognition as well as performance discussion on the reviewed methods. Furthermore, major limitations that hamper the development of automated micro-expression recognition systems are also analyzed, followed by recommendations of possible future research directions.

Keywords: Micro-expression; Apex frame; Spotting; Emotion Recognition; Deep Learning

INTRODUCTION

Micro-expression (ME) is a set of subconscious facial expressions that manifest less than 0.5s, yet it carries crucial and sufficient cues to disclose the real emotion of a person. An automated approach to ME is valuable in commercial and safety sectors such as for the application of clinical diagnosis, police interrogation, and national security (Yan et al. 2014). However, the short duration of ME, coupled with the low intensity of facial changes has posed a real challenge in designing an automated detection and recognition system. Besides that, manual detection and recognition of ME require a great deal of time and effort, even for a highly skilled expert. Therefore, an automated ME recognition system is a necessary tool to reduce the time-latency in identifying the right emotion. In general, the general flow of such a system can be simplified as shown in Figure 1.

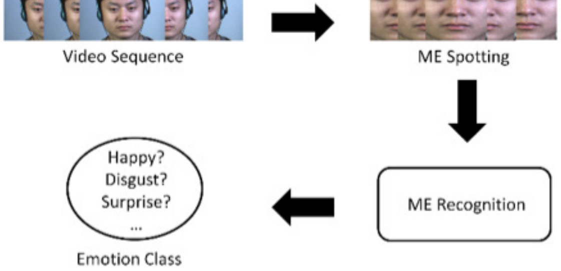

FIGURE 1. General flow of an automated micro expression recognition system

Nowadays, the training video samples containing MEs can be found publicly online, which will be used as input to the automated system. However, the available databases are usually constructed with posed ME events, whereby spontaneous databases are preferable for the development of an ME analyzer because they provide the real and true facial excitations. Usually, the input will pass through some pre-processing steps such as face detection to detect facial regions, face registration to align the detected face region, motion amplification, and temporal normalization. Viola-Jones classifier (Viola & Jones 2001), multi-task learning (X. Zhu & Ramanan 2012), and joint cascade method (D. Chen et al. 2014) are some of the popular methods used for facial detection. Even for some applications, the image will be transformed into a standardized form using the color constancy method (Zulkifley & Moran 2010). Similar to the traditional machine learning approach, some of the deep learning-based methods have also applied face detection module as the pre-requisite to the ME recognization module. Matsugu et al. (2003) utilized a convolutional neural network (CNN) and a rule-based algorithm as the basis for their face detection. A single CNN model has also been applied in Ranjan et al. (2019) work to classify an image, either it contains a face or not. On the other hand, CNNs have also achieved good performances for the application of face landmark localization (Deng et al. 2018), (Bian et al. 2018).

It is worth noting that deep learning methods have even been applied to the data pre-processing phase of ME analysis. In general, research in automated ME recognition involves two main parts: spotting and recognition. The former part concerns on the localization of the peak ME occurrence in a video, while the latter part focuses on the classification of the emotion categories based on MEs. Hence, this review paper will discuss the application of deep learning methods to both parts, spotting and recognition.

The second part of the paper will discuss the existing popular databases and their facial feature variations. After that, ME spotting and recognition will be discussed in sections 3 and 4, respectively. In section 5, the current challenges, general trends, and future work will be concisely reviewed. Finally, the main outcomes of ME spotting and recognition are concisely concluded in section 6.

DATABASE AND FEATURE

DATABASE

ME analysis systems conceived by the researchers need to be evaluated and validated on rigorous ME databases. Besides that, an automated ME system usually involves two modules, which are training and testing modules, whereby a leave-one-subject-out scheme is used to cross-validate system performance between various subjects. Hence, a good ME database is crucial in the development and evaluation of an emotion recognition system.

The posed ME database consists of deliberate expressions enacted by the subjects. These ME samples are collected by instructing the subjects to purposely produce the targeted emotion expression, which is a far cry from the unintentional nature of ME. As such, these datasets are heavily used during the early studies of ME and are not popular in the deep learning era. Table 1 is a summary of the posed datasets.

For ME analysis, spontaneous ME samples, which mimic closely the real-life expressions are needed. This selection is to ensure that ME analysis is capable to handle real-life ME challenges, including the preprocessing, spotting, and recognition stages. For example, during a police interrogation exercise, the subjects are expected to be far trickier and more professional in lying, as such they can mask their genuine emotion effectively. Table 2 summarized several popular public spontaneous ME datasets.

Spontaneous Micro-Expression Corpus (SMIC) consists of 51 samples of positive emotions, 70 negative samples, and 43 samples under surprise categories. The SMIC video samples were captured using a set of high-speed camera (HS), normal speed camera (VIS), and infrared camera (NIR). This dataset does not provide action units (AU) label and apex frame locations.

Chinese Academy of Sciences Micro-Expression (CASME) through CASME II, which is the improved version of CASME has analyzed 35 Chinese youths for the dataset development. The emotion categories of this dataset are happiness (33 samples), disgust (60 samples), surprise (25 samples), repression (27 samples), and others (102 samples). The major issue encountered by this dataset is the imbalanced data distribution between the emotions. Contrary to CASME II, SAMM has recruited 32 participants from 13 ethnicities to overcome the lack of ethnic diversity.

TABLE 1. Posed micro-expression datasets

Dataset	Reference	Subjects	Samples	No. Emotions	Emotions
Polikovsky	(Polikovsky et al. 2009)	11	42	7	Smile, Surprise, Anger, Sad, Disgust, Fear, Contempt
USF-HD	(Shreve et al. 2011)	-	100	4	Smile, Surprise, Anger, Sad
YorkDDT	(Warren et al. 2009)	9	18	2	Emotional, Non-emotional

TABLE 2. Spontaneous micro-expression datasets

Dataset	Reference	Subjects	Samples	No. Emotions	Emotions
SMIC	(Li et al. 2013)	16 (HS) 8 (VIS) 8 (NIR)	164 71 71	3 3 3	Positive, Negative, Surprise
CASME	(Yan et al. 2013)	35	195	7	Happiness, Sadness, Disgust, Surprise, Contempt, Fear, Repression or Tense
CASME II	(Yan et al. 2014)	35	247	5	Happiness, Disgust, Surprise, Repression, Others
CAS(ME)2	(Qu et al. 2018)	22	57	4	Positive, Negative, Surprise, Others
SAMM	(Davison et al. 2018)	32	159	7	Contempt, Disgust, Fear, Anger, Sadness, Happiness, Surprise

FEATURE DESCRIPTOR

ME features are the unique representation of the emotion extracted from the raw ME video samples. Early works on ME recognition were mostly done based on ME feature analysis, whereby many researchers believed that an improvement in a ME recognition system can be achieved by designing feature extractors that could best capture the nuances of the face. This assumption has led to the introduction of various features to optimally represent the facial characteristics.

Local Binary Pattern (LBP) is a type of appearance-based feature that is generated by calculating statistical features directly based on the pixel values. For LBP, the size relationship between a pixel point and its surrounding pixels is encoded in a binary form, which is then analyzed in a histogram-based representation. The generated histogram then will be the feature vector to represent the texture of an area of interest (Ojala et al. 1996). The general idea of LBP extraction is illustrated in figure 2.

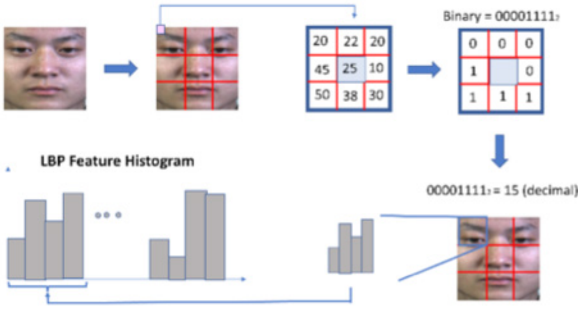

FIGURE 2. LBP feature extraction

The LBP representation is then improved by introducing a Local Binary Pattern on Three-Orthogonal Planes (LBP-TOP) (Zhao & Pietikäinen 2007). Graphically, a video sample can be considered as a cube in x, y and t dimensions as shown in figure 3. The xy, xt and yt ortogonal planes are first extracted, which are then stitched together to generate the final LBP-TOP features. This extended LBP feature extraction method was initially used to extract macro-expressions, which is then applied to ME as used in the studies by (Yan et al. 2014), (Li et al. 2013) and (Pfister et al. 2011).

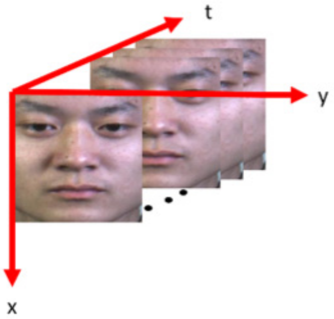

FIGURE 3. Micro-expression Video Three Orthogonal planes

Although LBP-TOP performs relatively well in the previous ME studies, but this method extracts the features in the form of high dimensionality, which is very difficult to down-scale to form sparse sampling representation. Therefore, several extensions of LBP were published such as Local Binary Pattern with Six Intersection Points (LBP-SIP) (Wang et al. 2015) and Local Binary Pattern with Mean Orthogonal Planes (LBP-MOP) (Wang et al. 2015).

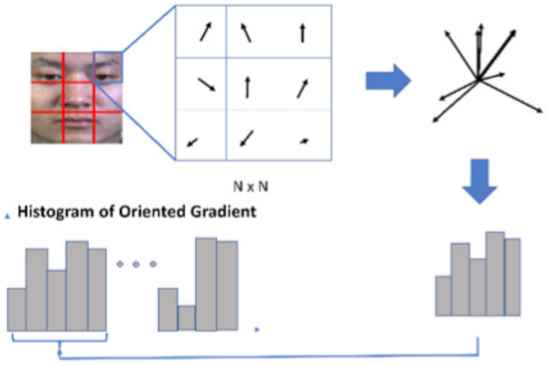

FIGURE 4. HOG feature extraction

Histogram of Oriented Gradients (HOG) is a gradient-based feature representation that is still heavily used today for object detection and image recognition (Li et al. 2018). HOG is known to be good in detecting corners and edges in an image, whereby the gradient value will increase significantly when there is a sharp change in intensity. The process of generating the HOG feature starts from computing the image gradient in both x and y directions, followed by constructing a histogram of gradients, which is then further processed to produce the final HOG descriptor vector as shown in figure 4.

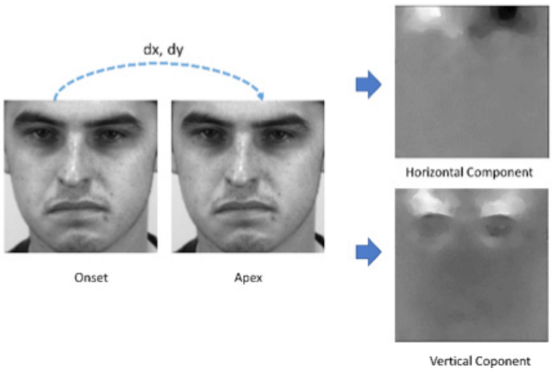

FIGURE 5. Optical flow between two frames

Contrary to the LBP feature descriptor, an optical flow descriptor does not consider directly the pixel values themselves, but rather the displacement values of certain feature points or areas of interest. Figure 5 illustrates the optical flow between the onset frame and apex frame of one set of SAMM ME data. This type of feature descriptor encodes the object movements through intensity changes of the image pixels. TV-L1(Zach et al. 2007) is one of the popular methods to calculate the optical flow approximation. While, Main Directional Mean Optical Flow (MDMO) as proposed by Liu et al. (Liu et al. 2016) considers the local static motion information and spatial location, which is found to produce better ME recognition results when it is compared to the LBP-TOP descriptor. In addition, Facial Dynamics Map (FDM) (Xu et al. 2017), Bi-Weighted Oriented Optical Flow (Bi-WOOF) (Liong et al. 2018)there is still plenty of room for improvement in terms of micro-expression recognition. Conventional feature extraction approaches for micro-expression video consider either the whole video sequence or a part of it, for representation. However, with the high-speed video capture of micro-expressions (100–200 fps, Histogram of Oriented Optical Flow (HOOF) (S. Zhang et al. 2017) were also used to approximate the optical flow for ME applications.

MICRO-EXPRESSION SPOTTING

Short Video Detection

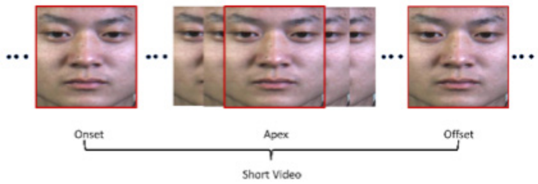

FIGURE 6. Micro-expression short video

ME spotting aims to locate the peak ME in a video, in other words, a short video of ME that was extracted from a raw long video. A short video of ME begins from the onset frame and ends once it reaches the offset frame as shown in Figure 6. An onset frame is the frame in which the expression starts to appear, while an offset frame is the frame in which the ME ends and reverts to the neutral expression. Among the frames in a short video, there is an apex frame that contains the greatest facial muscle movements. It plays an important role in the subsequent study of automated micro-expression analysis (Liong et al. 2016).

Shreve et al. (2009), (2011) detected and distinguished the occurrence of macro and micro expressions in a long video by calculating the strain magnitude, derived from the optical flow method. While Polikovsky et al. (2009) we present a novel approach for facial micro-expressions recognition in video sequences. First, 200 frame per second (fps), (2013) have utilized 3D-HOG descriptor as the input feature to locate the onset, apex, and offset frames of a video, captured using a high-speed camera with a frame rate of 200 frames per second. However, these studies were examined and validated using posed ME datasets: Shreve et al. (USF-HD dataset) and Polikovsky et al. (Polikovsky dataset). The obvious limitations of such datasets have been discussed in previous section.

Thereafter, many studies have shifted their focus to spontaneous databases, whereby the detection of such MEs is more challenging but more relevant to the real situation. Moilanen et al. (2014) have applied LBP- χ2 distance method for ME spotting tested on the CASME and SMIC datasets. The authors utilized appearance-based of LBP feature difference analysis, which then calculated the Chi-squared distance to measures the feature disparity, which are then further processed to detect a set of frames with ME movements that exceed a threshold value. This thresholding method has been adopted and modified by several later studies. Davison et al. have applied both, the HOG (Davison et al. 2015) and 3D-HOG (Davison et al. 2018) feature descriptors, which are then used in feature difference analysis coupled with thresholding of Chi-Squared (χ2) distances to spot the MEs. They have explored different databases of CASME and SMIC in (Moilanen et al. 2014), SAMM and CASME II in (Davison et al. 2018). Thus, it is not practical to directly compare which of these methods is the more effective one.

In (Patel et al. 2015), the onset and offset frames of the SMIC dataset were identified with the aid of optical flow vectors. The proposed algorithm was designed to capture the continuity information of the movement flows and directions. Li et al. (2018) have proposed the first automatic ME analysis system (MESR) that combines the thresholding value of feature contrast used in ME spotting and recognition tasks for a long spontaneous video (SMIC and CASME II). The feature extractors used in this paper are LBP and HOOF, of which the former method produced better performance. Besides, Wang et al. (2017) have also utilized the threshold technique for their proposed Main Directional Maximal Difference (MDMD) Analysis of optical flow features. The differences of the CAS(ME)2 frame features obtained using MDMD features are more pronounced than those of LBP features.

Apart from the popular thresholding technique, a random walk model was used in (Xia et al. 2016) to estimate the probability of the presence of ME in video frames. While Li et al. (2016) have used CNN for pre-processing the ME data. The ME detection is done based on a deep multi-task learning method with the HOOF input feature. In the later Micro-Expression Spotting Challenge 2019, the LTP machine learning method proposed by Li et al. (2019) have performed better on SAMM and CAS(ME)2 compared to the LBP- χ2 distance method.

APEX FRAME SPOTTING

Early studies on ME consider both the spatial image features and temporal timing features, thence the entire short video of the ME needs to be processed first before emotion can be recognized. However, some researchers only analyze the most crucial frame in the ME short video, which is the apex frame (Liong et al. 2018). It is a frame that lies between the onset frame (beginning) and the offset frame (ending) of a short video as depicted in Figure 3 It portrays the most expressive emotional state, thus the highest intensity of expression changes can be retrieved from this frame. Yet the spotting of apex frames can be a challenging task due to the short duration of the expression and subtle facial movement intensity.

The very first automated apex frame spotting research was designed by Yan et al. (2015), whereby the authors have employed two feature extractors, namely LBP and Constraint Local Models, pivoted on the assumption that an apex frame will have the largest feature differences among the subsequent frames. Liong et al. (2015) have argued that the maximum feature variations do not necessarily correspond to the apex frame. In addition to this, Liong et al. have pointed out two flaws in Yan et al.'s work that affects the practicability of the results. The performance analysis was deemed to be incorrectly done because the average frame distance between the spotted apex frames and the ground truth was not based on the absolute mean data. Besides that, they have also argued that the validation was only performed on about 20% of the video samples in the CASME II dataset.

Then, Liong et al. published an improved version of their apex spotting method in (Liong et al. 2015). An extra feature descriptor, the optical strain was added as part of the input features to the first apex spotting network. A divide-and-conquer strategy applied to the region of interest was suggested to locate the occurrence of the apex frame. After that, Liong et al. (2016) have further extended their work to recognize ME from a set of long videos, also based on a single spotted apex frame. The spotting task is done by the novel eye masking approach to exclude the irrelevant ME movements, which is then further processed by an optical strain feature descriptor. Ma et al. (2017) have proposed a method to automatically spot the apex frame using Region Histogram of Oriented Optical Flow (RHOOF) feature. The proposed RHOOF can reflect the changes in the facial movements for the video samples taken from CASME and CASME II datasets.

Zhang et al. (2018) then combined a deep learning approach and feature matrix processing method for the application of apex frame spotting. A new CNN network namely spotting micro-expression convolutional network (SMEConvNet) was designed to extract the relevant features. Although the deep neural network is less optimal for the application with a medium-size dataset, specifically the publicly available ME datasets, SMEConvNet that consists of four pairs of convolution and pooling (Conv-Pool) layers have managed to achieve 22.36 average frames difference on the long video input. The number of Conv-Pool layers in the network will heavily affect the performance of the deep learning network. A lesser number of Conv-Pool pairs will limit the network capability in extracting high-level features, while a larger number of layers will possibly cause an overfitting problem, which will subsequently cause worse performance. Table 3 summarized the performance of the reviewed apex spotting studies. The CASME-II-RAW database used in the studies is based on the raw long video of ME, which is a more challenging dataset compared to the short video.

The effectiveness evaluation of the apex frame spotting is usually measured using the Mean Absolute Error (MAE). MAE measures the average difference (in terms of frames) of the spotted apex frame compared to the ground truth label of the apex frame. The metric will produce a small MAE value for a well-performed apex spotting algorithm.

TABLE 3. Studies in Apex frame spotting

Paper	Method	Database	Best Result
Yan et al. (2015)	LBP CLM	CASME II	Mean: 0.31 Mean: 1.02
Liong et al. (2015)	LBP CLM OS	CASME II	MAE: 13.55 MAE: 17.21 MAE: 14.43
Liong et al. (2016)	LBP OS	CASME II-RAW	MAE: 55.26 MAE: 27.21
Ma et al. (2017)	RHOOF	CASME CASME II	MAE: 3.60 MAE: 10.97
Zhang et al. (2018)	SMEConvNet	CASME II-RAW	MAE: 22.36

MICRO-EXPRESSION RECOGNITION

TRADITIONAL MACHINE LEARNING APPROACHES

As mentioned in section Feature Descriptor, much of the early works in ME recognition are based on the traditional machine learning approach, which is highly dependent on the handcrafted feature extraction methods. The sub-optimal features are then passed to the traditional classifiers such as Support Vector Machines (SVM) (Suykens & Vandewalle 1999), Random Forest (RF) (Breiman 2001), Naive Bayes (MA et al. 2014), Multiple Kernel Learning (MKL) (Varma & Ray 2007) and Local Vector Space Model (LVSM) (Vu Thanh et al. 2015), to classify the features, which is then subsequently organized into a set of predetermined categories.

The most representative study would be the publication by Pfister et al. (2011). This paper can be considered as the pioneering work on automated ME recognition. The authors have utilized LBP-TOP extractor to extract ME features with the aid of MKL and Temporal Interpolation Model (TIM), which has achieved 71.4% accuracy tested on the early version of SMIC database that propelled the LBP-TOP to be the baseline comparison model in many subsequent works on ME analysis. Some of the publications that have analyzed hand-crafted ME features are (Yan et al. 2014), (Li et al. 2013), and (Liong et al. 2015). Among them, the work in (Liong et al. 2018) has used the SVM coupled with the Bi-WOOF feature for the analysis of Apex frame and Onset frame only. On this basis, the study has achieved considerably good results that outperform the other methods with F1-scores of 0.62 (SMIC-HS) and 0.61 (CASME II). This paper has also shown that the information contained in the Apex frame is able to represent the entire ME video, which has inspired the subsequent studies to focus solely on the relationship between Apex and Onset frames for ME analysis. Table 4 summarizes related ME recognition studies utilizing traditional machine learning approaches.

DEEP LEARNING APPROACHES

Rapid advances in the development of deep neural networks have influenced the progress in ME recognition, whereby there is a less and lesser method designed based on the hand-crafted features. Several researchers have proposed and implemented optimal feature learning using a deep learning approach that eliminates the need of extracting the features manually. As a result, the CNNs methodology has been applied to both spotting and recognition of ME systems. Figure 7 compares the traditional machine learning approach which requires the aid of handcrafted features and classifier with the deep learning approach.

Several famous CNN network structures such as VGG-M (Zulkifley & Trigoni 2018), ResNet (He et al. 2016), AlexNet (Krizhevsky et al. 2012), GoogleNet (Inception) (Szegedy et al. 2015), and VGGNet (Simonyan & Zisserman 2015) have been applied to the ME recognition. Compared to the tedious hand-crafted feature extraction in the traditional setting, deep learning networks extract a set of optimal features through deep recursive training. The general structure of the deep network often involves feature extraction through several layers of CNN, followed by fully connected layers for the classification task. The final layer is connected to several neurons that depend on the number of classes, whereby the probability distribution of the sample belonging to each class is obtained through the softmax activation function.

A concise summary of some published deep learning-based ME recognition studies can be found in Table 5. The table shows that ME recognition system that is based on the deep learning paradigm started around 2016 by Kim et al. (2016). The authors have combined CNN and long short-term memory (LSTM) to encode and process the spatial information of the following patterns: start, start to apex, apex, apex to end, and end.

Another interesting ME recognition study is the Off-ApexNet that was published by Gan et al. (2019), which has outperformed the other ME studies with the highest accuracy of 88.28% and F1-score of 0.8697, tested on the CASME II dataset. The overall procedure of their method started with locating the apex frame index using a divide-and-conquer strategy on the region of interest to attain the optical flow information of the apex frame and reference frame. Then, the optical flow features are fed into a pre-designed CNN model for further feature enhancement as well as expression classification. The evaluation of Off-ApexNet has been done on SMIC, CASME II, and SAMM databases. The recognition accuracy on the SMIC database is the lowest, indicating that there is still room for improvement in automatic apex frame spotting.

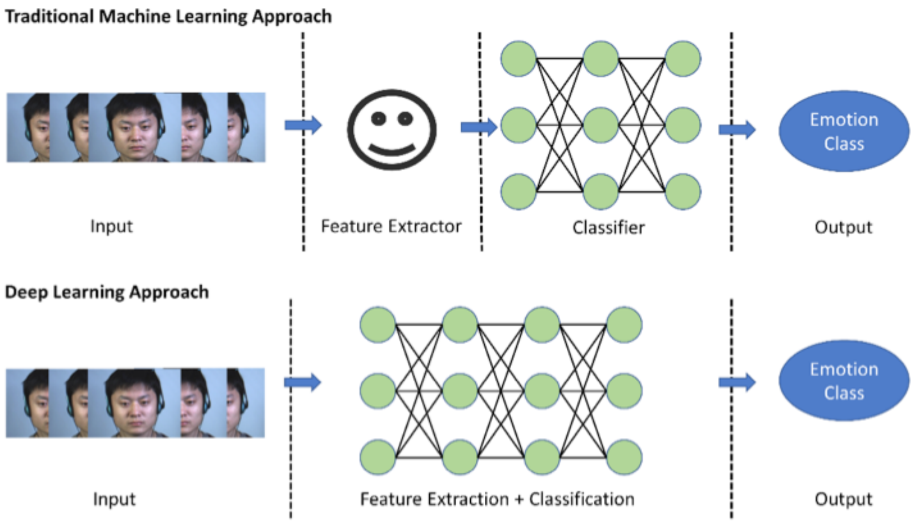

FIGURE 7. Traditional machine learning and deep learning approach

TABLE 4. ME recognition using traditional machine learning approaches

Feature types	Classifier	Reference	Database	Best Result
LBP-TOP	SVM, MKL, RF	Pfister et al. (2011)	Earlier version of SMIC	Acc: 71.4% (with MKL)
LBP-TOP	SVM	Li et al. (2013)	SMIC	Acc: 52.11% (on VIS)
LBP-TOP	SVM	Yan et al. (2014)	CASME II	Acc: 63.41%
OSW-LBP-TOP	SVM	Liong et al. (2014)	SMIC CASME II	Acc: 57.54% Acc: 66.40%
MDMO	SVM	Liu et al. (2016)	SMIC CASME CASME II	Acc: 80% Acc: 68.86% Acc: 67.37%
3D-HOG	Fuzzy	Chen et al. (2016)	CASME II (36 samples)	Acc: 86.67%
Bi-WOOF	SVM	Liong et al. (2016)	SMIC CASME II	F1: 0.62 (on HS) F1: .0.61
Bi-WOOF	SVM	Liong et al. (2016)	SMIC CASME II	Acc: 53.52% F1: 0.59
LBP-TOP, HOOF	RK-SVD	Zheng et al. (2016)	CASME CASME II	Acc: 69.04% Acc: 63.25%
LBP-TOP, Optical Flow	KNN, SVM, RF	Zhang et al. (2017)	CASME II	Acc: 62.50%
LBP-TOP, LBP-SIP, STLBP-IP	KGSL	Zong et al. (2018)	CASME II and SMIC	F1: 0.6125

TABLE 5. ME recognition using deep learning approaches

Method	Reference	Database	Best Result
CNN+LSTM	Kim et al. (2016)	CASME II	Acc: 60.98%
DTSCNN	Peng et al. (2017)	CASME and CASME II	Acc: 66.67%
ELRCN	Khor et al. (2018)	CASME II SAMM	F1: 0.5 F1: 0.409
ResNet	Wang et al.(2018)	SMIC CASME II SAMM	Acc: 49.4% Acc: 65.9% Acc: 48.5%
OFF-ApexNet	Gan et al. (2019)	SMIC CASME II SAMM	F1: 0.6709 F1: 0.8697 F1: 0.5423
3D-FCNN	Li et al. (2019)	SMIC CASME CASME II	Acc: 55.49% Acc:54.44% Acc: 59.11%
STSTNet	Liong et al. (2019)	SMIC CASME II SAMM	UF1: 0.6801 UF1: 0.8382 UF1: 0.6588
Apex-Time Network	Peng et al. (2019)	SMIC CASME II SAMM	UF1: 0.497 UF1: 0.523 UF1: 0.429
CapsuleNet	Van Quang et al. (2019)	SMIC CASME II SAMM	UF1: 0.5820 UF1: 0.7068 UF1: 0.6209
MER-RCNN	Xia et al. (2019)	SMIC CASME CASME II	Acc: 57.1% Acc: 63.2% Acc: 65.8%
Dual-Inception	Zhou et al. (2019)	SMIC CASME II SAMM	UF1: 0.6645 UF1: 0.8621 UF1: 0.5868
DSTICNN	Zhu et al. (2020)	SMIC CASME II	Acc: 85.93% Acc: 83.65%

PERFORMANCE ANALYSIS

The performance of a ME recognition system is usually assessed using Leave-One-Subject-Out (LOSO) cross-validation, as such the test will cover a wide range of emotions per subject.

TABLE 6. ME recognition confusion matrix

		Predicted Class	
		Yes	No
Actual Class	Yes	True Positive ()	False Negative ()
	No	False Positive ()	True Negative ()

The confusion matrix as shown in Table 6 is also used to supplement the classification performance metrics, apart from accuracy (Acc) and F1-score. Accuracy quantifies the ratio of correctly predicted observations with respect to the total observations. Although accuracy is a great metric, it can produce a biased result if the distribution between the sample size of true and false categories is greatly imbalanced, and hence will not reflect the real performance of the model. For such cases, F1-score is the better option to measure the performance of an imbalanced dataset distribution, which is a norm for ME analysis, whereby the subjects cannot invoke all the targeted emotions. This is because F1-score considers both the Precision and Recall metrics. Precision is defined as the measure of correctly identified positive cases from all the predicted positive cases, while Recall measures the ratio of correctly identified positive cases to all the actual positive cases. These performance metrics can be calculated as shown below:

$$Precision = \frac{T_P}{T_P + F_P} \quad (1)$$

$$Recall = \frac{T_P}{T_P + F_N} \quad (2)$$

$$F1\ score = \frac{2x(RecallxPrecision)}{Recall + Precision} \quad (3)$$

DISCUSSION AND FUTURE WORK

Automatic recognition of ME is now a popular researcher field, especially in the digital era, where an accurate measurement of the user emotion is crucial for targeted marketing purposes. In contrast to the early studies in this field, the number of available spontaneous ME databases has increased significantly, which will greatly facilitate and benefit the development of automated ME systems. However, the existing databases still have a few shortcomings that include uneven sampling distribution between the emotion categories and the homogeneity of the subject's ethnicity. For example, the CASME II database as discussed in this paper only considers Chinese youth as the subject, whereby the emotion invoked by them might be different compared to the other nationalities. Besides that, it is also inefficient for the largest sample of ME to belong to the "other" category that accounts for 64% of the total data, which is in contrast to the "surprise" category with only 15% of the sample. This imbalance between emotions is mainly due to the difficulty of capturing the targeted ME, but this is also an indicator that there is still sizable room for improvement in the development of ME analysis. The inadequacy of the ME database will hamper the rapid research and development of automated ME recognition systems. Thus, increasing the number of ME video samples is an unavoidable challenge that needs to be addressed immediately.

On the other hand, our previous discussion has also shown that a lot of researchers in the ME recognition system have gradually moved towards the deep learning-based paradigm. From the laborious and time-consuming process of manually spotting and recognition approach to the utilization of standard machine learning approach, deep learning has been implemented in the most important parts of the automated ME recognition system, which are spotting and recognition tasks. This is undoubtedly a great advancement in the ME analysis study. Deep learning has been well established in many other fields (Abdani & Zulkifley 2019), whereby its strength in terms of accuracy, efficiency, and timesaving have been proven in various ME recognition systems. For example, the previously reviewed Off-ApexNet which utilizes CNN with optical flow input has returned better apex frame spotting. Besides, the research direction of ME spotting and recognition should aim towards more practical applications, which consequently are in line with the aims of real-time ME recognition systems. The application of deep learning methodology in ME analysis has reduced the tedious and time-consuming efforts needed in manual detection and analysis.

However, the development of automatic ME recognition systems should not be limited to basic deep learning techniques. Although much research on automated ME recognition systems has been done but still a lot of state-of-art methods have only been evaluated on limited ME samples, and thus not optimal for real-life applications. Another important issue is that the development of ME spotting is far less researched compared to ME recognition. Li et al. (2018) have argued that the lack of precise ME spotting methods has significantly reduced the ME recognition system accuracy. ME spotting is not a new topic, but over the years, there is still no prominent breakthrough due to the lack of spontaneous ME databases, even the available databases are not challenging enough in mimicking real-life applications. Sizable ME samples are needed to explore more possibilities in ME analysis design so that accuracy can be improved. One possible research direction is the application of data augmentation to synthetically create additional data, which may help to overcome the problem of ME samples shortage.

CONCLUSION

In brief, this article collates and discusses the past and present development of automated ME analysis systems. In the early studies, the development is more skewed towards the standard machine learning approach, while the present ones are more skewed towards the deep learning approach. It is also worth to note that the improvement in ME spotting accuracy will directly improve the recognition performance of the ME analysis systems. Lastly, the limitations of current ME recognition systems and few future recommendations are also mentioned concisely.

ACKNOWLEDGEMENT

The authors would like to acknowledge funding from Universiti Kebangsaan Malaysia (Geran Universiti Penyelidikan: GUP-2019-008) and Ministry of Higher Education Malaysia (Fundamental Research Grant Scheme: FRGS/1/2019/ICT02/UKM/02/1).

DECLARATION OF COMPETING INTEREST

None

REFERENCES

Abdani, S. R. & Zulkifley, M. A. 2019. DenseNet with Spatial Pyramid Pooling for Industrial Oil Palm Plantation Detection. *Proceedings of the 2019 International Conference on Mechatronics, Robotics and Systems Engineering, MoRSE 2019,* 134–138.

Bian, P., Xie, Z. & Jin, Y. 2018. Multi-task feature learning-based improved supervised descent method for facial landmark detection. *Signal, Image and Video Processing* 12(1): 17–24.

Breiman, L. 2001. Random forests. *Machine Learning* 45(1): 5–32.

Chen, D., Ren, S., Wei, Y., Cao, X. & Sun, J. 2014. Joint cascade face detection and alignment. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), LNCS*, 109–122.

Chen, M., Ma, H. T., Li, J. & Wang, H. 2016. Emotion recognition using fixed length micro-expressions sequence and weighting method. *IEEE International Conference on Real-Time Computing and Robotics, RCAR 2016,* 427–430.

Davison, A. K., Lansley, C., Costen, N., Tan, K. & Yap, M. H. 2018. SAMM: A Spontaneous Micro-Facial Movement Dataset. *IEEE Transactions on Affective Computing* 9(1): 116–129.

Davison, A. K., Yap, M. H. & Lansley, C. 2015. Micro-Facial Movement Detection Using Individualised Baselines and Histogram-Based Descriptors. *Proceedings - 2015 IEEE International Conference on Systems, Man, and Cybernetics, SMC 2015*, 1864–1869.

Davison, A., Merghani, W., Lansley, C., Ng, C. C. & Yap, M. H. 2018. Objective micro-facial movement detection using FACS-Based regions and baseline evaluation. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*, 642–649.

Deng, W., Fang, Y., Xu, Z. & Hu, J. 2018. Facial landmark localization by enhanced convolutional neural network. *Neurocomputing* 273: 222–229. doi:10.1016/j.neucom.2017.07.052

Gan, Y. S., Liong, S. T., Yau, W. C., Huang, Y. C. & Tan, L. K. 2019. OFF-ApexNet on micro-expression recognition system. *Signal Processing: Image Communication* 74: 129–139.

He, K., Zhang, X., Ren, S. & Sun, J. 2016. Deep residual learning for image recognition. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*: 770–778.

Khor, H. Q., See, J., Phan, R. C. W. & Lin, W. 2018. Enriched long-term recurrent convolutional network for facial micro-expression recognition. *Proceedings - 13th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2018*: 667–674.

Kim, D. H., Baddar, W. J. & Ro, Y. M. 2016. Micro-expression recognition with expression-state constrained spatio-temporal feature representations. *Proceedings of the 2016 ACM Multimedia Conference*: 382–386.

Krizhevsky, A., Sutskever, I. & Hinton, G. E. 2012. ImageNet Classification with Deep Convolutional Neural Networks. Advances in Neural Processing Systems 25 (NIPS 2012).

Li, J, Wang, Y., See, J. & Liu, W. 2019. Micro-expression recognition based on 3D flow convolutional neural network. *Pattern Analysis and Applications* 22(4): 1331–1339.

Li, J, Soladie, C., Sguier, R., Wang, S. & Yap, M. H. 2019. Spotting Micro-Expressions on Long Videos Sequences. *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*.

Li, X, Hong, X., Moilanen, A., Huang, X., Pfister, T., Zhao, G. & Pietikainen, M. 2018. Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods. *IEEE Transactions on Affective Computing* 9(4): 563–577.

Li, X, Pfister, T., Huang, X., Zhao, G. & Pietikainen, M. 2013. A Spontaneous Micro-expression Database: Inducement, collection and baseline. *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013*.

Li, X, Yu, J. & Zhan, S. 2016. Spontaneous facial micro-expression detection based on deep learning. *International Conference on Signal Processing Proceedings, ICSP*: 1130–1134.

Liong, S.-T., Gan, Y. S., See, J., Khor, H.-Q. & Huang, Y.-C. 2019. Shallow Triple Stream Three-dimensional CNN (STSTNet) for Micro-expression Recognition. *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*.

Liong, S. T., See, J., Phan, R. C. W., Le Ngo, A. C., Oh, Y. H. & Wong, K. S. 2015. Subtle expression recognition using optical strain weighted features. *Lecture Notes in Computer Science, LNCS* 9009: 644–657.

Liong, S. T., See, J., Wong, K. & Phan, R. C. W. 2015. Automatic micro-expression recognition from long video using a single spotted apex. *Lecture Notes in Computer Science*, *LNCS* 10117: 345–360.

Liong, S. T., See, J., Wong, K. S., Le Ngo, A. C., Oh, Y. H. & Phan, R. 2015. Automatic apex frame spotting in micro-expression database. *Proceedings - 3rd IAPR Asian Conference on Pattern Recognition, ACPR 2015*: 665–669.

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

Liu, Y. J., Zhang, J. K., Yan, W. J., Wang, S. J., Zhao, G. & Fu, X. 2016. A Main Directional Mean Optical Flow Feature for Spontaneous Micro-Expression Recognition. *IEEE Transactions on Affective Computing* 7(4): 299–310.

Ma, H., An, G., Wu, S. & Yang, F. 2017. A Region Histogram of Oriented Optical Flow (RHOOF) feature for apex frame spotting in micro-expression. *2017 International Symposium on Intelligent Signal Processing and Communication Systems, ISPACS 2017 - Proceedings*: 281–286.

Matsugu, M., Mori, K., Mitari, Y. & Kaneda, Y. 2003. Subject independent facial expression recognition with robust face detection using a convolutional neural network. *Neural Networks*, 16: 555–559.

Moilanen, A., Zhao, G. & Pietikäinen, M. 2014. Spotting rapid facial movements from videos using appearance-based feature difference analysis. *Proceedings - International Conference on Pattern Recognition,* 1722–1727.

Ojala, T., Pietikäinen, M. & Harwood, D. 1996. A comparative study of texture measures with classification based on feature distributions. *Pattern Recognition* 29(1): 51–59.

Patel, D., Zhao, G. & Pietikäinen, M. 2015. Spatiotemporal integration of optical flow vectors for micro-expression detection. *Lecture Notes in Computer Science, LNCS* 9386: 369–380.

Peng, M., Wang, C., Bi, T., Chen, T., Zhou, X. & shi, Y. 2019. A Novel Apex-Time Network for Cross-Dataset Micro-Expression Recognition. 8th International Conference on Affective and Intelligent Interaction, 1-6.

Peng, M., Wang, C., Chen, T., Liu, G. & Fu, X. 2017. Dual temporal scale convolutional neural network for micro-expression recognition. *Frontiers in Psychology* 1745.

Pfister, T., Li, X., Zhao, G. & Pietikäinen, M. 2011. Recognising spontaneous facial micro-expressions. *Proceedings of the IEEE International Conference on Computer Vision* 1449–1456.

Polikovsky, S., Kameda, Y. & Ohta, Y. 2009. Facial micro-expressions recognition using high speed camera and 3D-Gradient descriptor. *IET Seminar Digest*, 2009.

Polikovsky, S., Kameda, Y. & Ohta, Y. 2013. Facial micro-expression detection in hi-speed video based on facial action coding system (FACS). *IEICE Transactions on Information and Systems* E96-D(1): 81–92.

Qu, F., Wang, S. J., Yan, W. J., Li, H., Wu, S. & Fu, X. 2018. CAS(ME)2): A Database for Spontaneous Macro-Expression and Micro-Expression Spotting and Recognition. *IEEE Transactions on Affective Computing* 9(4): 424–436.

Ranjan, R., Patel, V. M. & Chellappa, R. 2019. HyperFace: A Deep Multi-Task Learning Framework for Face Detection, Landmark Localization, Pose Estimation, and Gender Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 41(1): 121–135.

Shreve, M., Godavarthy, S., Goldgof, D. & Sarkar, S. 2011. Macro- and micro-expression spotting in long videos using spatio-temporal strain. *IEEE International Conference on Automatic Face and Gesture Recognition and Workshops, FG 2011*: 51–56.

Shreve, M., Godavarthy, S., Manohar, V., Goldgof, D. & Sarkar, S. 2009. Towards macro- and micro-expression spotting in video using strain patterns. *Workshop on Applications of Computer Vision, WACV 2009*.

Simonyan, K. & Zisserman, A. 2015. *Very Deep Convolutional Networks for Large-Scale Image Recognition*. The 3rd International Conference on Learning Representations.

Suykens, J. A. K. & Vandewalle, J. 1999. Least squares support vector machine classifiers. *Neural Processing Letters* 9(3): 293–300.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., et al. 2015. Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 1–9.

Van Quang, N., Chun, J. & Tokuyama, T. 2019. CapsuleNet for micro-expression recognition. *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*.

Varma, M. & Ray, D. 2007. Learning the discriminative power-invariance trade-off. *Proceedings of the IEEE International Conference on Computer Vision*.

Viola, P. & Jones, M. 2001. Rapid Object Detection using a Boosted Cascade of Simple Features. *Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition.*

Vu Thanh, N., Trong Le, T., Tuan Dinh, L. & Huynh Nguyen, K. H. 2015. Local Classification Model with Vector Space for Multi-Class Text Classification. *IEEE RIVF International Conference on Computing & Communication Technologies*.

Wang, C., Peng, M., Bi, T. & Chen, T. 2018. Micro-Attention for Micro-Expression recognition. *Neurocomputing* 410: 354–362.

Wang, S. J., Wu, S., Qian, X., Li, J. & Fu, X. 2017. A main directional maximal difference analysis for spotting facial movements from long-term videos. *Neurocomputing* 230: 382–389.

Wang, Y., See, J., Phan, R. C. W. & Oh, Y. H. 2015. Efficient spatio-temporal local binary patterns for spontaneous facial micro-expression recognition. *PLoS ONE* 10(5): e0124674.

Wang, Y., See, J., Raphael, R. & Oh, Y. H. 2015. LBP with six intersection points: Reducing redundant information in LBP-TOP for micro-expression recognition. *Lecture Notes in Computer Science, LNCS* 9003, 525–537.

Warren, G., Schertler, E. & Bull, P. 2009. Detecting deception from emotional and unemotional cues. *Journal of Nonverbal Behavior* 33(1): 59–69.

Xia, Z., Feng, X., Hong, X. & Zhao, G. 2019. Spontaneous facial micro-expression recognition via deep convolutional network. *8th International Conference on Image Processing Theory, Tools and Applications, IPTA 2018 Proceedings.*

Xia, Z., Feng, X., Peng, J., Peng, X. & Zhao, G. 2016. Spontaneous micro-expression spotting via geometric deformation modeling. *Computer Vision and Image Understanding* 147: 87–94.

Xu, F., Zhang, J. & Wang, J. Z. 2017. Microexpression Identification and Categorization Using a Facial Dynamics Map. *IEEE Transactions on Affective Computing* 8(2): 254–267.

Yan, W.-J., Wu, Q., Liu, Y.-J., Wang, S.-J. & Fu, X. 2013. CASME Database: A Dataset of Spontaneous Micro-Expressions Collected From Neutralized Faces. *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition, FG 2013, 1-7.*

Yan, W. J., Li, X., Wang, S. J., Zhao, G., Liu, Y. J., Chen, Y. H. & Fu, X. 2014. CASME II: An improved spontaneous micro-expression database and the baseline evaluation. *PLoS ONE* 9(1).

Yan, W. J., Wang, S. J., Chen, Y. H., Zhao, G. & Fu, X. 2015. Quantifying micro-expressions with constraint local model and local binary pattern. *Lecture Notes in Computer Science* 8925: 296–305.

Zach, C., Pock, T. & Bischof, H. 2007. A duality based approach for realtime TV-L1 optical flow. *Lecture Notes in Computer Science LNCS* 4713: 214–223.

Zhang, S., Feng, B., Chen, Z. & Huang, X. 2017. Micro-expression recognition by aggregating local spatio-temporal patterns. *Lecture Notes in Computer Science, LNCS* 10132: 638–648.

Zhang, Z., Chen, T., Meng, H., Liu, G. & Fu, X. 2018. SMEConvNet: A Convolutional Neural Network for Spotting Spontaneous Facial Micro-Expression from Long Videos. *IEEE Access* 6: 71143–71151.

Zhao, G. & Pietikäinen, M. 2007. Dynamic texture recognition using local binary patterns with an application to facial expressions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 29(6): 915–928.

Zheng, H., Geng, X. & Yang, Z. 2016. A relaxed K-SVD algorithm for spontaneous micro-expression recognition. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)* 9810 LNCS: 692–699.

Zhou, L., Mao, Q. & Xue, L. 2019. Dual-inception network for cross-database micro-expression recognition. *Proceedings - 14th IEEE International Conference on Automatic Face and Gesture Recognition, FG 2019*.

Zhu, W. & Chen, Y. 2020. Micro-expression recognition convolutional network based on dual-stream temporal-domain information interaction. *Proceedings - 2020 13th International Symposium on Computational Intelligence and Design, ISCID 2020,* 396–400.

Zhu, X. & Ramanan, D. 2012. Face detection, pose estimation, and landmark localization in the wild. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, 2879–2886.

Zong, Y., Huang, X., Zheng, W., Cui, Z. & Zhao, G. 2018. Learning from hierarchical spatiotemporal descriptors for micro-expression recognition. *IEEE Transactions on Multimedia* 20(11): 3160–3172.

Zulkifley, M. A. & Moran, B. 2010. Enhancement of robust foreground detection through masked GreyWorld and color co-occurrence approach. *Proceedings - 2010 3rd IEEE International Conference on Computer Science and Information Technology, ICCSIT 2010* 4: 131–136.

Zulkifley, M. A., Mustafa, M. M., Hussain, A., Mustapha, A., & Ramli, S. 2014. Robust identification of polyethylene terephthalate (PET) plastics through Bayesian decision. *PloS one* 9(12).

Zulkifley, M. A. & Trigoni, N. 2018. Multiple-model fully convolutional neural networks for single object tracking on thermal infrared video. *IEEE Access* 6: 42790–42799