Knowledge-Based Systems 212 (2021) 106566

Contents lists available at ScienceDirect

Knowledge-Based Systems

journal homepage: www.elsevier.com/locate/knosys

LGAttNet: Automatic micro-expression detection using dual-stream local and global attentions

Madhumita A. Takalkar [a,*,1], Selvarajah Thuseethan [b,1], Sutharshan Rajasegarar [b], Zenon Chaczko [a], Min Xu [a,*], John Yearwood [b]

[a] University of Technology Sydney, Ultimo, NSW 2007, Australia
[b] Deakin University, Geelong, VIC 3220, Australia

ARTICLE INFO

Article history:
Received 15 June 2020
Received in revised form 15 September 2020
Accepted 23 October 2020
Available online 2 November 2020

Keywords:
Micro-expression detection
Dual-stream network
2D-CNN
Attention network

ABSTRACT

Research in the field of micro-expressions has gained significance in recent years. Many researchers have concentrated on classifying micro-expressions in different discrete emotion classes, while detecting the presence of micro-expression in the video frames is considered as a pre-requisite step in the recognition process. Hence, there is a need to introduce more advanced detection models for micro-expressions. In order to address this, we propose a dual attention network based micro-expression detection architecture called LGAttNet. LGAttNet is one of the first to utilize a dual attention network grouped with 2-dimensional convolutional neural network to perform frame-wise automatic micro-expression detection. This method divides the feature extraction and enhancement task into two different convolutional neural network modules; sparse module and feature enhancement module. One of the key modules in our approach is the attention network which extracts local and global facial features, namely local attention module and global attention module. The attention mechanism adopts the human characteristic of focusing on the specific regions of micro-movements, which enables the LGAttNet to concentrate on particular facial regions along with the full facial features to identify the micro-expressions in the frames. Experiments performed on widely used publicly available databases demonstrate the robustness and superiority of our LGAttNet when compared to state-of-the-art approaches.

1. Introduction

One of the most natural ways for individuals to communicate their feelings and thoughts is through facial expressions. Perhaps the correct perception of feelings from facial expressions is the most significant social activity that people, as social beings, perform [1]. Not all feelings, though, will be reflected on the face. Given the attempts to conceal, studies have discovered that real feelings are always leaked. Such leaked feelings typically manifest as micro-expressions [2]. Micro-expression is a brief facial expression, lasts for the overall duration of less than 500 ms and the onset duration of less than 260 ms [3]. Typically, it happens in circumstances of high stakes, particularly for people who win or risks something valuable [2]. The precise identification of such micro-expressions provides a tremendous ability for those with face-to-face communications expertise, including health care professionals, psychotherapists, educators and law enforcement officers because of their involuntary nature [4,5]. Moreover, recognizing micro-expressions is regarded as one of the most accurate tools for identifying deceit, owing to the near association between micro-expressions and deception [6].

Similar to macro-expressions, the micro-expression research has gained popularity in recent years due to its significance in many application domains, starting from psychology to medical [7]. Automatic macro-expression detection and recognition can be accomplished with the advent of technology in real-time and is effectively implemented in the industry since macro-expressions are clear to recognize and last for 500 ms to 4 s [8]. In contrast to macro-expression, a micro-expression is harder to recognize due to its subtle presence in the facial regions, which makes the detection and recognition using naked eye challenging to accomplish in real-time. According to Ekman,[2] sometimes the micro-expressions can be quicker than usual, and even occur for less than 40 ms. Ekman further indicated that the detection and

[*] Corresponding authors.
E-mail addresses: madhumita.a.takalkar@student.uts.edu.au (M.A. Takalkar), tselvarajah@deakin.edu.au (S. Thuseethan), srajas@deakin.edu.au (S. Rajasegarar), zenon.chaczko@uts.edu.au (Z. Chaczko), min.xu@uts.edu.au (M. Xu), john.yearwood@deakin.edu.au (J. Yearwood).

[1] Both authors have contributed equally.

[2] https://www.paulekman.com/resources/micro-expressions/.

https://doi.org/10.1016/j.knosys.2020.106566
0950-7051/© 2020 Elsevier B.V. All rights reserved.

recognition of micro-expressions are much more challenging in comparison to spotting. Detecting micro-expressions has to be performed using the images while neglecting the temporal information connected to micro-expressions. Moreover, much of the literature related to micro-expression focuses on spotting [9,10] and recognition [11–13], and a little research has been dedicated to the detection, which is the foundation of this study.

The continuing technological innovation in the field of computer vision and machine learning helps in boosting the recognition efficiency, and alleviates the issues related to micro-expression detection to a certain extend. As spontaneous micro-expressions can often be seen in real life and reveal better affective knowledge about humans, this work concentrates on the issue of detecting spontaneous micro-expressions from video frames effectively. Thus far, many micro-expression detection approaches analyze the disparity in the features between the first frame and the other frames in a time span [14–16]. In contrast, this work aims to detect the micro-expressions from the spatial features that can be extracted from a single video frame.

In the case of micro-expressions, it is interesting to note that most of the clues originate from a few facial regions such as the mouth and eyes. Ideally, this suggests that the machine learning models must concentrate only on the relevant facial areas and be less responsive to the other facial regions. The predictions made by convolutional neural networks (CNNs) are based on the posterior probability functions, whereas the professionals typically render judgments that can be clarified more clearly depending on the selective local facial regions of interests (RoIs). Similar to this human behavior, the attention mechanism can also concentrate on specific regions of images. Much research lately aims to incorporate attention mechanisms with deep networks [17–19]. Through deep learning, the attention model directly simulates the human brain's attention mechanism. In earlier research, the importance of incorporating the attention system has been thoroughly discussed [9,10,20–24].

In this paper, a deep learning based framework for micro-expression detection is proposed, which applies an attention mechanism to concentrate on the salient parts of the face. The overall architecture of the proposed framework is presented in Fig. 1 and further elaborated in Section 3. The following are the novel and key contributions of this paper:

1. An attention driven detection mechanism, called LGAttNet, to identify the frame-wise micro-expression is proposed. According to the best of our knowledge, this is the first approach to use a dual attention network for building a micro-expression detection framework. LGAttNet is designed to be an automatic micro-expression detection model that focuses on particular facial regions with specific information related to micro-expressions on top of the full face information.
2. The attention networks in LGAttNet are structured as dual-stream local and global attention blocks. The local attention stream of the architecture focuses on the RoIs that exist only within local facial areas for associated micro-muscle movements, whereas the global attention stream considers the full face, establishing a relation between the local facial RoIs. Further, to extract the global and local attention maps, a combination of deep and shallow networks is used instead of using a single very deep CNN, wherein the deep network provides a sparse representation of the features that are subsequently passed on to a traditional shallow CNN with the sigmoid function.
3. The performance of the proposed LGAttNet micro-expression detection framework is evaluated on publicly available and widely used CASME, CASME II, CAS(ME)^2 and SAMM databases using a leave-one-subject-out (LOSO) cross-validation. A comprehensive ablative study to manifest that the idea of building a dual-stream network using local and global attention networks achieves improvement in the micro-expression detection accuracy is conducted. A cross-database analysis is also performed to verify the efficiency of the proposed architecture. Furthermore, a comparison of the LGAttNet with state-of-the-art approaches is performed to demonstrate that the LGAttNet performs remarkably well in detecting the micro-expressions in video frames.

The remaining parts of this paper are structured in the following order. Section 2 summarizes the related studies performing binary classification between micro-expression and non micro-expression video frames. Section 3 introduces our framework, called LGAttNet, and its supporting components. Experimentation performed on the model along with the outcomes is presented in Section 4. Finally, the paper concludes in Section 5 with future research directions.

2. Related research

In spite of the fact that automatic micro-expression detection and recognition are not broadly analyzed in contrast to macro-expression study, a number of works have addressed this problem with the recent advances in computer vision. For micro-expression analysis, micro-expression detection is a crucial and essential pre-processing phase in defining a corresponding series of frames from a given long video containing micro-expressions. Given that micro-expression is an uncontrolled facial expression, the micro-expression detection study has been conducted on the publicly available spontaneous micro-expression repositories. It is difficult to distinguish concise facial gestures from neutral faces, especially in real life videos and to prevent false alarms triggered by global facial actions, speaking and occlusions.

Many approaches proposed for micro-expression detection primarily focus on assessing the discrepancy between their own features, which indicates the disparity in the time window from the first and the other frames. Since the span of the spontaneous micro-expression is relatively short, only a few frames are available in a video that reveal the micro-expressions rendering the detection of spontaneous micro-expression extremely hard. The techniques used in the literature to detect these micro-expressions are broad, including optical flow [25], Local Binary Patterns (LBP) [26], Histogram of Oriented Gradients (HOG) [27] and integral projection [28].

Davison et al. [29,30] implemented "individualized baselines" determined by taking the participant's neutral video sequence and using the Chi-square distance to achieve the initial features for baseline sequence. Lu et al. [28] introduced a low-computing cost approach focusing on differences in the integral projection (IP) of sequential micro-expression frames for detection. Li et al. [14] published a spontaneous micro-expression database (SMIC), offering the benchmark results for detection and recognition of micro-expressions. The authors further observed that not every subject exhibited micro-expressions while capturing samples for SMIC. Micro-expression detection was carried out in a two-class classification process by differentiating the micro-expression clip from a randomly selected non micro-expression clips. Throughout this analysis, the researchers applied an active shape model (ASM), which normalizes and monitors all faces to focus on spatial feature variations and local binary pattern-three orthogonal planes (LBP-TOP) for extraction of features. In order to facilitate feature extraction, the temporal interpolation model (TIM) was implemented to adjust frame numbers, and the

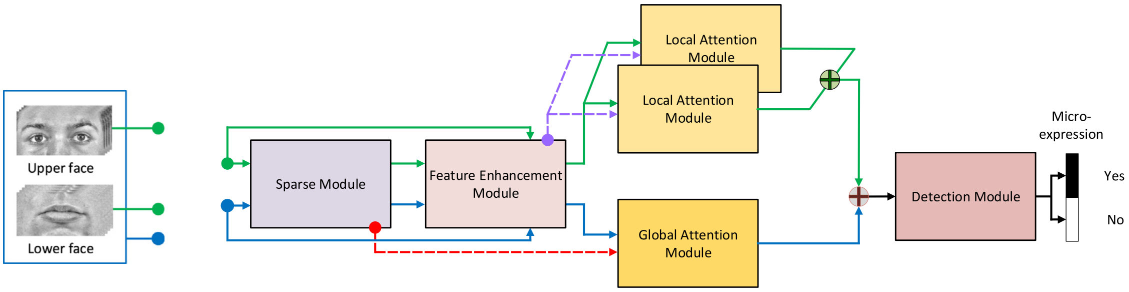

Fig. 1. The overall architecture of the proposed Local and Global Attention Network (LGAttNet) model. LGAttNet takes three inputs: upper face, lower face and entire face. The flows of the upper and lower face images are indicated by the green color arrows, while the flow of the entire face is indicated by blue arrows. Both dotted lines indicate the attention inputs to global and local attention modules. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

classification was performed using the support vector machine (SVM).

Huang et al. [31] implemented spatio-temporal completed local quantization patterns (STCLQP) to perform the detection by extracting sign, magnitude and orientation as features. The movement magnitude across frames has been used by Borza et al. [15] with simple absolute frame variations, along with an Adaboost algorithm to identify micro-expression frames. Deep learning based techniques have been used for micro-expression detection in the past. Li et al. [32] introduced a deep multi-task approach with HOOF analysis for micro-expression detection, using CNN for pre-processing micro-expression data to recognize the location of the facial landmarks and split the facial area into regions of interest. The sliding window based technique proposed by Borza et al. [16] preserves the present frame, past and future frame at equivalent intervals, with the discrepancy between these being given to the CNN that categorizes the period as micro-expression or non micro-expression. In addition, Zhang et al. [33] have implemented a deep learning technique to identify micro-expressions from longer videos for the first time. A novel CNN based architecture, namely SMEConvNet (spotting micro-expression convolutional network), was developed for the extraction of the features from video clips. For a long video apex frame spotting, the feature matrix processing method using the sliding window technique was also proposed to consider micro-expression characteristics in order to search for the apex frame.

Besides the existing deep learning systems, the network can also focus on certain facial regions by incorporating the attention mechanism to the micro-expression recognition architecture. Attention enhances the representation of interests besides simply showing where to concentrate. Fernandez et al. [19] introduced a CNN-based end-to-end approach utilizing attention methodology to address facial expression recognition problems for representation and classification jointly. Likewise, in 2D + 3D facial expression recognition, Jiao et al. [20] suggested enhanced facial attention-based convolutional neural network (FA-CNN). The facial attention mechanism allows the network to automatically identify the discriminative regions without dense landmark annotations from multi-modal expressions. Wang et al. [9] designed a novel attention model, namely the micro-attention to help in emphasizing the facial RoIs. For precise micro-expression recognition, Yang et al. [10] applied visual attention to developing an attention-based CNN network called MERTA. Although attention has been actively applied for face recognition [23], facial expression recognition [19,20] and recently also for micro-expression recognition [9,10], it has not yet been considered for facial micro-expression detection.

Nonetheless, the aforementioned studies are correlated with micro-expression spotting and attention networks for facial expression recognition while the actual work for detecting micro-expression using attention network has never been studied. The occurrence of micro-expression in small sections of the face and the insufficient size of available repositories hinder the precision of recognition. In this work, we propose a mechanism to incorporate attention network for micro-expression detection with the available amount of data samples.

3. LGAttNet detection model description

LGAttNet model is the first to utilize the sparse representation and attention mechanism for micro-expression detection. The first phase of our architecture is the pre-processing activity. Subsequently, the pre-processed image is utilized in three different ways. The image is initially divided into two parts, where the first part focuses on the eye regions and the second part focuses on the mouth region. Further, the full facial image is also used for processing. The output of the final module represents the results in different metrics, predicting the existence of micro-expressions.

Our proposed model, dual-stream LGAttNet, comprises five modules: Sparse module (SM), Feature Enhancement module (FEM), Local Attention module (LAM), Global Attention module (GAM) and Detection module (DM) as presented in Fig. 1.

The input to the LGAttNet is a pre-processed video frame. The input pre-processed face image is divided into two subparts: upper face, focusing mainly on the eyes and eye-brows muscles of the face, and lower face, which concentrates on the mouth section of the face. Here, the image size of both upper and lower faces are made identical to the entire face by converting the rest of area to a black background. Each of these face sections, as well as the whole face region are given as input to the SM, which is a deep CNN network without fully connected and classifier layers. Deep features from SM along with the respective input image are fed into the FEM, a shallow network with a Concatenation and Sigmoid functions.

Each LAM shown in the diagram is a dedicated attention module for upper and lower face parts, respectively. There are two input values given to both LAMs. The first feature vector is the output vector of the FEM Sigmoid layer for respective upper and lower face regions after concatenating with the corresponding input facial RoI-based features. The second input feature vector is the feature vector before the input image concatenation, which is the output from the last Convolutional layer for each face image part. Apart from focusing on local face parts, SM and FEM also processes full face image and feeds the feature vectors from SM and FEM as input to the GAM, the third attention module. The inclusion of GAM helps to preserve the relationship between the upper and lower face. In the end, the output vectors from upper and lower face LAM and GAM are concatenated and given to the DM, which is a traditional deep neural network with a classifier layer to predict the existence of the micro-expression in the input video frame.

Following subsections explain these modules in detail.

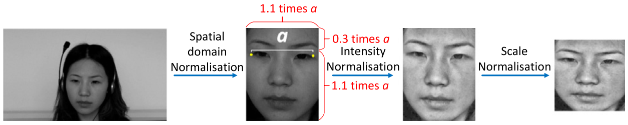

Fig. 2. Pre-processing steps: *Spatial domain normalization* is achieved through the difference "*a*" between facial feature points 37 and 46 of the active appearance model (AAM). The *intensity* and *scale* normalizations are performed subsequently.

3.1. Pre-processing

The input video frames are translated to grayscale during the pre-processing stage so that the cross-database variations between the video frames can be minimized. The descriptor involves two essential pre-processing steps: (a) data augmentation or producing synthetic samples and (b) normalization. During the data augmentation phase, a series of synthesized frames are created in large amounts to enhance the number of video frames, particularly for training with a deep learning model, which usually requires a larger database. Random noise is applied to the center of the eyes and nose regions of the face using a 2D Gaussian distribution to generate synthetic frames, following the method stated in [34]. This is applied to each of the images included in the micro-expression database. A micro-expression detection module is trained using individual frames.

Inspired by the work in [35], a set of normalization operations are then carried out in a sequence. Initially, a region of interest (ROI) is chosen in the process of spatial normalization for feature extraction, which excludes the insignificant areas of the video frames. This process discards the background details as well as some facial areas like ears, chin and forehead, since these regions represent no particular information regarding micro-expressions. The distance denoted as "*a*" in Fig. 2, between the active appearance model (AAM) points 37 and 46 is used to crop the facial region. Secondly, using Contrast Limited Adaptive Equalization (CLAHE) [36] approach, an intensity normalization step is implemented on every video frame to minimize the feature vector variance. One benefit of CLAHE is that the histogram segment that goes beyond the clip boundary between all histogram bins is redistributed instead of merely deleting it. A Rayleigh distribution with a cap of 0.01 and α value of 1 is chosen for this function. Thirdly, the video frames, in the scale normalization phase, are downsized by linear interpolation to 128 × 128 pixels. Scale normalization makes it possible for the same facial feature points to co-exist roughly at the same position in different video frames.

3.2. LGAttNet components

The SM and FEM are two deep 2D-CNN components of the proposed architecture. Further, a carefully crafted attention network consisting of 2D-convolutional and sigmoid layers is placed alongside SE and FEM composed of deep CNN, while eliminating the final softmax layer to give room for the Detection module.

3.2.1. Sparse Module (SM)

The image processing applications expect the input data to be represented in as few components as possible for faster processing. The sparse coding technique is useful for solving the classification problems where specific dictionaries for respective classes are built and the input is processed to classify the dictionary corresponding to the sparsest representation. In recent years, sparsity is applied to a variety of image processing and computer vision applications, such as image classification [37,38], image de-noising [39], compression [40], in-painting [40], object tracking [41] and other applications.

In general, SM is implemented using a deep convolutional neural network without the fully connected and classification layers. In similar way, our SM architecture is implemented using a deep convolutional neural network consisting of seven convolutional layers with one additional last layer, i.e. eight, convolutional layer to extract the attention map of the input image as depicted in Fig. 3. Note that SM takes the 3-channel input images of upper face, lower face and entire face in 128 × 128 dimension. The input to the SM is an image I and the output is the feature map M_{sm} processed from the input image.

$$M_{sm} = f_{sm}(I) \tag{1}$$

where, f_{sm} is the function of SM. The SM outputs three feature maps for three different images namely; upper face RoIs, lower face RoIs and the whole face image. The feature maps of the whole face image generated by SM are fed to FEM, and as an attention input to GAM.

3.2.2. Feature Enhancement Module (FEM)

As presented in Fig. 4, the FEM is a shallow 2D convolutional neural network with two traditional convolutional blocks. The input to this FEM is the output feature vector of the sparse representation module. Similar to the SM, FEM also consists of an additional convolutional layer to process the attention mapping using the extracted features from SM and FEM together. The respective sparse representation of upper and lower face regions is again convolved to obtain a feature vector, and extracted from the third convolutional layer. The output of the last convolutional layer preserves the sparsity of the image by not losing much high end representation. This extracted feature vector is passed on to the next relevant LAM.

Contrasting from SM, the FEM further extends its performance by incorporating concatenation and sigmoidal functions. As the name suggests, this module enhances the collected features by using the concatenation function to integrate the feature vector extracted after the third (or last) convolutional layer of FEM and the feature mapping of the input image (which includes upper, lower and entire face image) to generate a new enhanced feature vector. This operation helps to enhance the representation ability of the proposed micro-expression detection model. The next step is to pass the enhanced feature vector to the sigmoid function. Basically, the sigmoid function is used because it ranges between 0 to 1. Therefore, it is used for the models where the output is the probability value for prediction. As the probability range is only within 0 and 1, thus sigmoid is the choice for the model.

$$\begin{aligned} I_{fem} &= f_{fem}(M_{sm}) \\ M_{fem} &= S(Concat(I_{fem}, I)) \end{aligned} \tag{2}$$

where, f_{fem}, $Concat$ and S are the FEM, concatenation and sigmoid functions, respectively and feature map M_{fem}.

The prediction probability for upper and lower facial region images forms the input to the LAMs and that of the full face image is given as input to GAM.

3.2.3. Local Attention Module (LAM) and Global Attention Module (GAM)

The LGAttNet is built using three attention blocks: two LAMs and one GAM. The implementation of both LAM and GAM is transformed as follows.

$$\begin{aligned} A_{gam} &= M_{fem} \times C(M_{sm}) \\ A_{lam} &= I_{fem} \times C(M_{fem}) \end{aligned} \tag{3}$$

where, C is the convolution layer function in both attention modules, and A_{gam}, A_{lam} are the attention feature maps for GAM and LAM, respectively. An element-wise multiplication operation

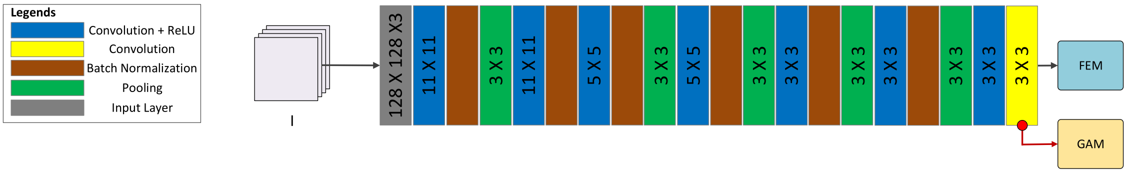

Fig. 3. Sparse module: 'I' is the input image given to the Sparse module, an eight-layer CNN including the Attention network. The attention feature map is extracted from the last convolutional layer and fed to the next modules—FEM and GAM.

is performed in both cases to obtain the attention feature maps. The two LAM processes the feature maps for upper face and lower face input images, respectively, whereas the GAM processes the feature vector for the full face. LAM focuses on the local facial regions, while GAM considers the complete face and preserves the relationship between the upper and the lower part of the face while displaying a micro-expression.

After constructing the attention feature maps, two feature level fusions are utilized, in order to get the resultant input to the detection module. To fuse the same size attention feature vectors of both LAMs, a mean operation is used. Subsequently, a concatenation operation is utilized to combine the resultant LAM based attention vector with the GAM based attention vector. The resultant feature map A_{res} is given by:

$$A_{res} = A_{gam} + \left(A_{lam}^{upper} + A_{lam}^{lower}\right) \qquad (4)$$

where, A_{lam}^{upper} and A_{lam}^{lower} are the attention feature maps produced for upper and lower parts of the face, respectively.

The LAM and GAM are composed of two components, as can be seen from Fig. 5; (1) 2D-CNN with one convolutional block and (2) a multiplication function. The architecture for LAM and GAM is the same, but the input given is different, and this is what makes each of the block function and process the input differently. All three attention blocks accept two input feature vectors. The first input to both the LAM is the sigmoid function probability output of the FEM for the upper and lower facial region images, respectively (solid green arrows towards LAMs in Fig. 1). This input feature vector given to respective attention blocks is then made to pass through a 2D-CNN with one convolutional block within the attention module. The second input is the feature representation extracted from the last convolutional layer of FEM (purple dashed arrows towards LAMs in Fig. 1). The convolved feature vector and the second input are then forwarded to the multiplication function to get a final representation of the respective upper and lower face regions.

Similarly, the sigmoid function output of FEM for full face image is passed to the Global Attention module (GAM) as a first input (solid blue arrow towards GAM in Fig. 1). The first input is convolved when given to GAM. The second input is the features collected from the last convolutional layer of SM (dashed red arrow towards GAM in Fig. 1). The output of CNN within GAM is then forwarded to the multiplication function along with the second input. Upon multiplication, a new feature representation is generated.

It should be noted that the second input for all the attention blocks is directly provided to the multiplication function, whereas the first input is convolved. The output vectors from all the attention blocks are added to form one vector and given to the Detection Module for the final detection of micro-expression.

3.2.4. Detection Module (DM)

The DM in LGAttNet consists of three fully connected (FC) layers in the size of 1024, 1024 and 512. In addition, a softmax classification layer is attached at last to perform the classification task. The estimation of micro-expression using DM is explained in Eq. (5).

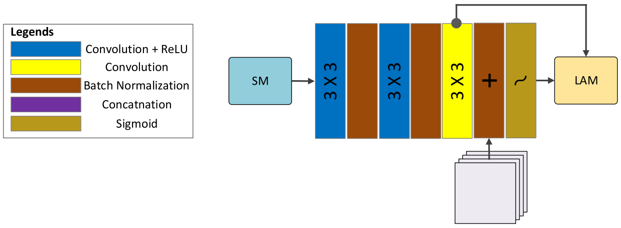

Fig. 4. Feature enhancement module: Input given to this shallow three-layer convolutional network is the attention feature map extracted from SM. The last convolutional layer extracts the attention feature map preserving the sparse but essential representations. Summation function generates a new enhanced feature map, and Sigmoid function outputs the prediction probability.

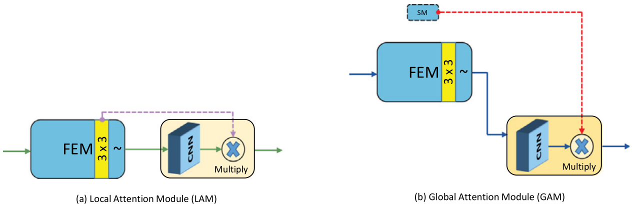
FEM
3 x 3
~
CNN
Multiply
(a) Local Attention Module (LAM)
SM
FEM
3 x 3
~
CNN
Multiply
(b) Global Attention Module (GAM)

Fig. 5. Local and global attention module: The first input fed to the one-layer convolutional attention network with a multiplication function, LAM and GAM, is the Sigmoidal probability extracted from FEM. The second input for LAM is the attention feature map from FEM and for GAM is the attention feature map from SM.

$$\hat{y} = f_{dm}(A_{res}) \qquad (5)$$

where, $\hat{y}$ is the prediction of an image sample.

3.3. Loss function

In order to train the proposed model, the degradation function is incorporated. Generally, the binary cross entropy (BCE) performs better in closed set classification tasks. Hence, the BEC is used in the proposed approach to estimate the classification loss, as given below.

$$\mathcal{L}_{cls} = \frac{1}{N}\sum_{i=0}^{N}(y \times \log \hat{y}_i) + (1-y) \times \log(1-\hat{y}_i) \qquad (6)$$

where, y and $\hat{y}$ are the micro-expression label and the predicted value, respectively.

4. Experimental setup and results

The validation and efficiency of LGAttNet are verified by testing the model on some of the publicly available benchmark micro-expression databases. Apart from the model testing, the effectuality of implementing local and global attention networks in a dual-stream pattern is demonstrated using an ablation study.

4.1. Datasets used

The extensive experiments are conducted on widely used micro-expression databases, including the Chinese Academy of Sciences micro-Expression datasets: CASME [42] and CASME II [43], the Chinese academy of sciences macro and micro-expressions (CAS(ME)^2) [44] and the spontaneous actions and micro-movements (SAMM) [29]. The details of these spontaneous micro-expression databases used in the experiments are given below.

Introduced by the Chinese academy of sciences, CASME database [42], is one of the spontaneous micro-expression database widely used. CASME comprises of two subsets A and B totaling up to 195 micro-expression samples collected from 19 participants and recorded at 60 fps. The video clips in dataset A were captured in natural light with a resolution of 1280 × 720 pixels. In dataset B, the video samples were recorded at 640 × 480 pixel resolution under LED lighting. Each sample was tagged with onset, apex and offset frames, action units (AUs) labeled and emotions correctly identified by psychologists.

As an extension of the original CASME [42] dataset, CASME II [43] was introduced with some additional samples. There are 247 micro-expressions newly coded with FACS captured from 26 participants under a high temporal resolution of 200 fps and a 280 × 340 pixels spatial resolution on the facial region to examine muscle movements in greater detail. Every video session is a short clip for several seconds, which has onset, apex and offset frames marked for micro-expressions and facial action coding system (FACS) and emotion types are annotated.

The CAS(ME)^2 database is a collection of long macro- and micro-expression videos. There are 87 long video samples of both the expression types with an average duration of 148 s collected from 22 participants in Part A of CAS(ME)^2 database and Part B is further divided into two sub-parts with 300 macro-expression samples and 57 micro-expression samples [44]. The database is annotated with the onset, apex and offset frame index for these expressions. Furthermore, onset and offset time for the eye blinks are also marked.

In the SAMM database [29], there are altogether 32 participants, recording seven video samples, each with an average duration of 35.5 s. The spotting task is emphasized on 79 videos where every video includes one or several facial micro-movements summing up to 159 micro-movements. As the ground truth, the onset, apex, and offset frame indices of micro-movements are given wherein the micro-movement duration lasts between the onset and the offset frame. All micro-movements in this database are labeled. The identified frames can, therefore, signify not only micro-expression but also other facial movements, including blinks of the eyes.

4.2. Experimental setup and parameters

The model implementation is done using an open-source platform, Tensorflow. The model is trained and tested on a GPU Server with NVIDIA GeForce GTX 1080 Titan processor. LGAttNet uses SGD optimization technique with an initial learning rate of 0.001 and L2 normalization is implemented to prevent overfitting. The input dimensions of the proposed LGAttNet model is 128 × 128. The other parameter values are weight decay of 0.0005 and a momentum of 0.9. The model training is executed for 100 epochs and early stopping is employed.

The results obtained for the conducted experiments are reported using accuracy, area under curve (AUC), F1-score, recall and precision.

Table 1
Micro-expression detection outcome using various performance metrics.

Metrics	Database			
	CASME	CASME II	CAS(ME)^2	SAMM
Precision	0.948	0.944	0.850	0.851
Recall	0.915	0.940	0.885	0.890
F1-score	0.931	0.942	0.867	0.870
Accuracy	0.932	0.942	0.865	0.867
TPR	0.915	0.940	0.885	0.890
FPR	0.050	0.055	0.155	0.155
AUC	0.931	0.912	0.923	0.846

4.3. Outcomes and analysis

The observations are drawn by performing analysis on four publicly available databases, namely CASME, CASME II, CAS(ME)^2 and SAMM. The inputs to the network are images or video frames from micro-expression databases. In order to confirm the effectiveness of the LGAttNet model, five metrics i.e., accuracy, precision, recall, F1-score, and area under curve (AUC), are chosen as evaluation metrics for binary classification.

4.3.1. Outcomes

The evaluation of the network is conducted using the Leave-One-Subject-Out Cross-Validation (LOSOCV) technique, i.e., one subject is selected that is not used for the training process, and the network is evaluated on this unseen subject. The training and testing set consists of two classes, wherein one class has micro-expression video frames, and the other class has neutral face video frames (non-ME frames). The macro-expression samples are eliminated from our experiments. As this is a subject independent evaluation, one subject is entirely left out of the training process.

Table 1 demonstrates the experiment outcomes for all the performance metrics. The table depicts that the sparsely represented multi-attention micro-expression detection architecture is capable of achieving significantly high, which is in the range of 87% to 94%, detection accuracy for different databases. As can be seen, the recognition accuracies obtained on CAS(ME)^2 and SAMM datasets are lower compared to other micro-expression datasets. It is clear that the samples in SAMM datasets contain wide variations, such as included subjects from different nationalities.

An additional experiment is performed to demonstrate that the LGAttNet works equally well on a sequence of video frames for detecting micro-expression frames. In this experiment, during the testing phase, the input given to the trained LGAttNet is a series of images from a video. LGAttNet then processes the video frame-wise and classifies the individual frame as ME or non-ME frame, as shown Fig. 6. The graph illustrates that the LGAttNet detects the micro-expression frames (green line in the graph) from the sequence of video frames in alignment with the ground-truth (red line in the graph). Therefore, it can be seen that the LGAttNet model trained on images is capable of accurately predicting the micro-expression frames from videos.

4.3.2. Ablative analysis

Table 2 shows the performance of ablative analysis on all the databases using accuracy metric. For undertaking these analyses, the model is modified and evaluated by including or removing each of the LAM and/or GAM component in the architecture. From the table, it can be observed that the LGAttNet performs well when it includes LAM and GAM on all the databases. The next evaluation is performed by removing the GAM module, where the results reveal a drop by 3% − 5%. The probable reason is that the network is unable to find the relation between the two individual

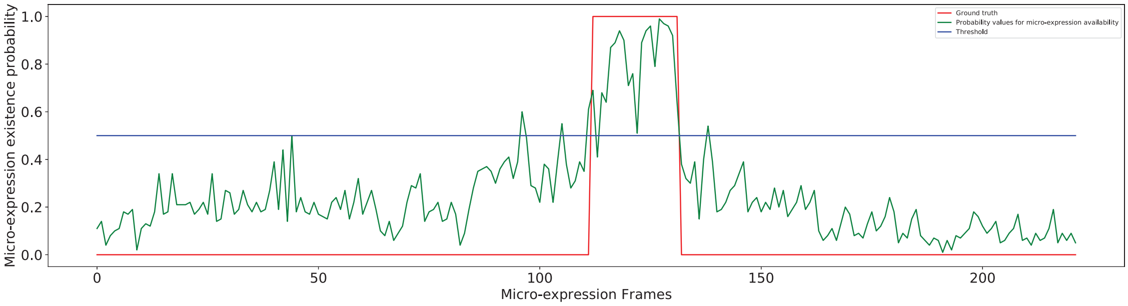

Fig. 6. LGAttNet tested on a micro-expression sequence from CASME database. The green line indicates the generated probability values for the existence of micro-expression. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Table 2
Recognition accuracy for ablative evaluation with and without (w/o) different modules of LGAttNet.

	CASME	CASME II	CAS(ME)^2	SAMM
LGAttNet	0.933	0.943	0.865	0.868
LGAttNet w/o GAM	0.892	0.915	0.812	0.822
LGAttNet w/o LAM	0.835	0.878	0.785	0.788
LGAttNet w/o GAM and LAM	0.735	0.782	0.692	0.728

local feature maps extracted from upper and lower face regions when the global attention module (GAM) is not present in the architecture, resulting in performance degradation.

However, it can be observed that removing LAM from the model has a significant effect on the results. The performance is seen to deteriorate more when the LAM is removed, while the GAM is included in the system. This shows that acquiring the local level features from the facial regions assists in interpreting micro-expression in the input image. Finally, removing both GAM and LAM decreases the results by more than 20%. These results demonstrate the importance of our proposed LAM and GAM modules for correctly detecting micro-expressions.

In addition to the ablative analysis, as seen in Table 2, Fig. 7 visualizes the influence of utilizing attention mechanism in LGAttNet for Disgust micro-expression. The purpose of incorporating attention mechanism in LGAttNet is to focus the attention towards specific facial regions to identify the presence of micro-movements and classify the input image as ME or non-ME frame. Similar to ablative analysis, Fig. 7 displays the attention mapping for LGAttNet with and without LAM or GAM or both. It can be seen from Fig. 7(b), which is the activation map for LGAttNet without both LAM and GAM, that the activation map is scattered all over the facial region. The model was unable to highlight the specific facial regions of movements without the attention mechanisms. However, Fig. 7(c) is an implementation of LGAttNet with LAMs and without GAM. We can observe that the LAM serves its purpose of concentrating on the specific regions of the face, which is eyes and mouth in this case. Fig. 7(d) is the activation map for LGAttNet with GAM and without LAMs, which highlights the central region of the face, including eyes, nose and upper lip. This indicates that unlike LAM, which processes upper and lower face individually, GAM has to process the full face. Hence, it is forced to process the facial regions which do not carry any movement and highlighting regions of non-interest. Finally, Fig. 7(e) is our complete model of LGAttNet with LAM and GAM. The activation map illustrates that the LAM that processes upper and lower face separately can predict the region of interest which carry a micro-expression and GAM which processes the full face correlates to the LAM features to precisely mark the facial regions with motion, eliminating any unnecessary details, classifying the input frame correctly as ME frame.

Table 3
Cross-database micro-expression detection network trained on CASME and tested on other databases.

Metrics	Database		
	CASME II	CAS(ME)^2	SAMM
Precision	0.866	0.829	0.768
Recall	0.845	0.850	0.745
F1-score	0.855	0.839	0.756
Accuracy	0.857	0.837	0.760
TPR	0.845	0.850	0.745
FPR	0.130	0.175	0.225
AUC	0.861	0.802	0.715

4.3.3. *Cross-database analysis*

The model effectiveness is also shown by conducting experiments on cross databases. The purpose of performing the cross-database evaluation is to justify that even though this architecture is a supervised learning model, it is capable of detecting micro-expressions from the images of altogether different databases. Cross-database micro-expression detection is where the training and testing samples come from two different micro-expression databases collected by different cameras or under different environments. The databases used in these experiments also have diversity in ethnic backgrounds of the participants. For instance, the SAMM database contains the samples of the subjects from 13 different nationalities. This type of validation offers a good way to mimic the scenarios the micro-expression detection system would encounter in reality. Therefore, it is worthy to investigate this more carefully.

In the experiments, four micro-expression databases are employed. In cross-database, the network is trained using the samples from only one database, and the testing is performed using the samples from the remaining three databases. For instance, in Table 3, the network is trained using CASME database and CASME II, CAS(ME)^2 and SAMM databases are utilized for testing. Similarly, the testing is carried out on the remaining databases, as can be seen from the results in Tables 4–6. The detection accuracy range for CASME, CASME II and CAS(ME)2 databases range between 0.82 and 0.86, i.e. only 3–4% difference whereas if compared to SAMM database the detection accuracy varies between 0.73–0.76 (Tables 3–5). Furthermore, the detection accuracy, in Table 6, is ranging from 0.67 to 0.70. The results show a noticeable descend in the performance when LGAttNet is trained on SAMM. The observations of the cross-database evaluations are elaborated in Section 4.4. These results demonstrate that our proposed architecture is generic, and capable of performing subject independent detection.

4.3.4. *Comparison with state-of-the-art*

Table 7 compares the proposed technique with the existing approaches. It should be noted that the experimental configurations for the methods compared may differ. The results for the

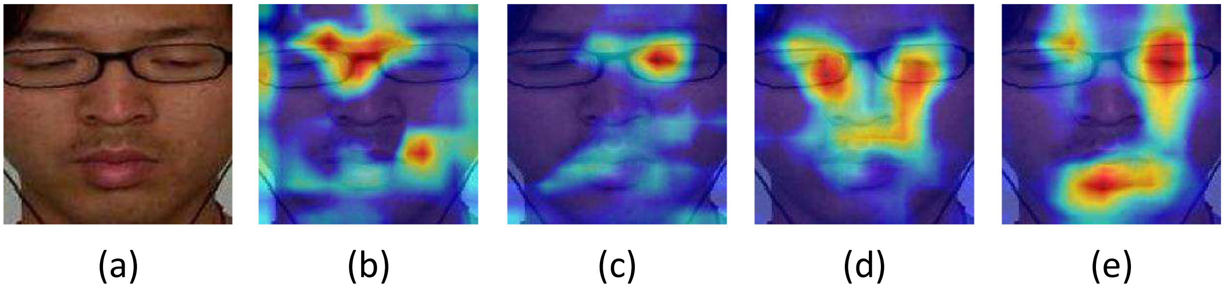

Fig. 7. LGAttNet attention visualization on a Disgust sample from CASME dataset. (a) Original image; (b) LGAttNet without LAM and GAM; (c) LGAttNet without GAM; (d) LGAttNet without LAM; (e) LGAttNet with LAM and GAM.

Table 4
Cross-database micro-expression detection network trained on CASME II and tested on other databases.

Metrics	Database		
	CASME	CAS(ME)^2	SAMM
Precision	0.891	0.810	0.733
Recall	0.825	0.835	0.730
F1-score	0.857	0.822	0.731
Accuracy	0.862	0.820	0.732
TPR	0.825	0.835	0.730
FPR	0.100	0.195	0.730
AUC	0.891	0.781	0.705

Table 5
Cross-database micro-expression detection network trained on CAS(ME)^2 and tested on other databases.

Metrics	Database		
	CASME	CASME II	SAMM
Precision	0.848	0.834	0.742
Recall	0.810	0.855	0.720
F1-score	0.828	0.844	0.730
Accuracy	0.832	0.842	0.735
TPR	0.810	0.855	0.720
FPR	0.145	0.170	0.250
AUC	0.849	0.842	0.710

Table 6
Cross-database micro-expression detection network trained on SAMM and tested on other databases.

Metrics	Database		
	CASME	CASME II	CAS(ME)^2
Precision	0.668	0.678	0.715
Recall	0.685	0.645	0.680
F1-score	0.676	0.661	0.697
Accuracy	0.672	0.670	0.703
TPR	0.685	0.645	0.680
FPR	0.340	0.305	0.272
AUC	0.631	0.665	0.697

Table 7
Comparison with existing state-of-the-art micro-expression detection methods.

Database	Method	Performance	Accuracy
CASME	Feature difference (CNN) [16]	TPR=77.27%	–
	(CASME-A) LBP-χ^2 [26]	–	78.75%
	(CASME-A) LTP-ML [26]	–	77.90%
	(CASME-B) LBP-χ^2 [26]	–	82.92%
	(CASME-B) LTP-ML [26]	–	82.61%
	LGAttNet	**TPR = 91.5%**	**93.2%**
CASME II	Frame difference [15]	–	81.75%
	Frame difference [45]	–	86.95%
	LBP-χ^2 [26]	–	64.08%
	LTP-ML [26]	–	65.07%

	LBP [46]	TPR=70.0%
		FPR=13.5%
			–

	PLK+LSTM [47]	–	89.87%
	LGAttNet	**TPR = 94.0%** **FPR = 5.5%**	**94.2%**
CAS(ME)^2	LTP-ML [48]	F1-score=0.0055	–
	LBP [44]	AUC=0.5971	–
	LGAttNet	**F1-score = 0.867** **AUC = 0.923**	**86.5%**
SAMM	3D HOG-XY plane [27]	–	70.87%
	LBP-TOP-XY plane [27]	–	74.65%
	HOOF [27]	–	70.98%
	LTP-ML [48]	F1-score=0.0316	–
	PLK+LSTM (SAMM+CASME II) [47]	–	87.30%
	LGAttNet	**F1-score = 0.870**	**86.7%**

TPR = True Positive Rate; FPR = False Positive Rate [47] implements cross-database experiment (SAMM+CASME II).

existing approaches are taken directly from the respective research studies. Many studies have been tested only using CASME II and SMIC repositories to detect micro-expressions. LGAttNet is also trained and tested on CAS(ME)^2 and SAMM databases in addition to the commonly used databases for micro-expression detection. It can be found from the comparison table Table 7 that the LGAttNet achieves considerably higher detection accuracy.

4.4. Discussion

LGAttNet is built using attention networks that are made to focus on three different sections of an input facial image, making this the first attempt to use attention network for local as well as global attention mapping for a facial image. Unlike some other related studies, LGAttNet also stands out for achieving profoundly high micro-expression detection accuracy from video frames. There have been some studies [25,49] performed to recognize facial micro-express-ions from single apex frames. Taking motivation from these works, a detection model is constructed to identify micro-expression images from non-micro-expression images. The observations in Table 1 demonstrates that the LGAttNet is capable of detecting the existence of the micro-expression in the video frames. Usually, detection of micro-expressions is performed by taking feature differences between consecutive frames or comparing the first reference frame with the rest of the frames in a video using handcraft feature descriptors. In contrast, the LGAttNet uses a deep attention network that can concentrate on the local as well as global facial regions to track the feature difference to detect micro-expressions.

It can be contemplated from the results in Tables 3–6 that the databases having participants from a similar ethnic background, as in CASME, CASME II and CAS(ME)^2, display higher prediction accuracy when trained on one of these databases as compared to the other database (SAMM), which has participants from 13 different nationalities. Moreover, when trained on SAMM, the detection accuracies for CASME, CASME II and CAS(ME)^2 declines as the training database includes only three Chinese participants, which is contrary to the other databases. From this cross-database analysis, it can also be understood that the people from different ethnic backgrounds have their unique ways of hiding real emotions. Hence, it is not only the way of capturing these micro-expressions that affects the accuracies, but also the ethnicities of the participants' plays an important role.

5. Conclusion and future direction

In this work, a deep learning model is designed to focus on specific facial regions and establish a correlation between these regions and the whole facial area. The proposed model, namely LGAttNet is a micro-expression detection model that incorporates the attention network to converge the network processing towards selected regions of the face. The LGAttNet comprises a deep and a shallow CNN supported by local and global attention networks and an Artificial Neural Network (ANN) for binary classification. The local attention network processes partial facial parts, and the global attention network operates on the complete facial image. This model is an image-based supervised detection model with non-ME and ME classes.

Our model is the first to implement an attention based network for micro-expression detection. As compared to the available number of related state-of-the-art micro-expression detection works, LGAttNet model delivers exceedingly higher detection accuracy around more than 9%. This is because of the inclusion of the attention nets, since the micro-expression is more of a spatial feature and with the partitioning of the face, the attention nets insist the network to focus on selected facial regions. This behavior of LGAttNet can be observed from the ablative analysis, where on the removal of the local attention module (LAM), the detection accuracy is negatively affected and also on entirely removing the attention modules (LAM and GAM) the network accuracy drops significantly.

Cross-database evaluations are conducted to explain the robustness of the proposed network and to demonstrate that this model can be useful for real-time processing. We are currently working on extending this model, which is capable of being implemented on the video sequences by considering the temporal dimension that can benefit the real-time processing industry in the near future. In the future, the model would also be extended to spot micro-expressions from live-stream videos.

CRediT authorship contribution statement

Madhumita A. Takalkar: Data curation, Investigation, Methodology, Software, Writing - original draft. **Selvarajah Thuseethan:** Methodology, Software, Validation, Writing - review & editing. **Sutharshan Rajasegarar:** Conceptualization, Supervision, Writing - review & editing. **Zenon Chaczko:** Conceptualization, Supervision, Writing - review & editing. **Min Xu:** Conceptualization, Supervision, Writing - review & editing. **John Yearwood:** Conceptualization, Supervision, Writing - review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] P.M. Niedenthal, M. Brauer, Social functionality of human emotion, Annu. Rev. Psychol. 63 (2012) 259–285.

[2] P. Ekman, Lie catching and microexpressions, Phil. Decept. 1 (2) (2009) 5.

[3] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, X. Fu, How fast are the leaked facial expressions: The duration of micro-expressions, J. Nonverbal Behav. 37 (4) (2013) 217–230.

[4] J. Endres, A. Laidlaw, Micro-expression recognition training in medical students: a pilot study, BMC Med. Educ. 9 (1) (2009) 47.

[5] P.J. Marsh, M.J. Green, T.A. Russell, J. McGuire, A. Harris, M. Coltheart, Remediation of facial emotion recognition in schizophrenia: Functional predictors, generalizability, and durability, Amer. J. Psychiatr. Rehabil. 13 (2) (2010) 143–170.

[6] M.G. Frank, E. Svetieva, Microexpressions and deception, in: Understanding Facial Expressions in Communication, Springer, 2015, pp. 227–242.

[7] S. Thuseethan, S. Rajasegarar, J. Yearwood, Complex emotion profiling: an incremental active learning based approach with sparse annotations, IEEE Access 8 (2020) 147711–147727.

[8] F. De la Torre, W. Chu, X. Xiong, F. Vicente, X. Ding, J. Cohn, Intraface, in: 2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), Vol. 1, 2015, pp. 1–8, http://dx.doi.org/10.1109/FG.2015.7163082.

[9] C. Wang, M. Peng, T. Bi, T. Chen, Micro-attention for micro-expression recognition, 2018, arXiv preprint arXiv:1811.02360.

[10] B. Yang, J. Cheng, Y. Yang, B. Zhang, J. Li, Merta: micro-expression recognition with ternary attentions, Multimedia Tools Appl. (2019) 1–16.

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

[12] M.A. Takalkar, H. Zhang, M. Xu, Improving micro-expression recognition accuracy using twofold feature extraction, in: International Conference on Multimedia Modeling, Springer, 2019, pp. 652–664.

[13] M.A. Takalkar, M. Xu, Z. Chaczko, Manifold feature integration for micro-expression recognition, Multimedia Systems 26 (5) (2020) 535–551.

[14] X. Li, T. Pfister, X. Huang, G. Zhao, M. Pietikäinen, A spontaneous micro-expression database: Inducement, collection and baseline, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–6.

[15] D. Borza, R. Danescu, R. Itu, A. Darabant, High-speed video system for micro-expression detection and recognition, Sensors 17 (12) (2017) 2913.

[16] D. Borza, R. Itu, R. Danescu, Micro expression detection and recognition from high speed cameras using convolutional neural networks., in: VISIGRAPP (5: VISAPP), 2018, pp. 201–208.

[17] W. Sun, H. Zhao, Z. Jin, A visual attention based roi detection method for facial expression recognition, Neurocomputing 296 (2018) 12–22.

[18] S. Minaee, A. Abdolrashidi, Deep-emotion: Facial expression recognition using attentional convolutional network, 2019, arXiv preprint arXiv:1902.01019.

[19] P.D. Marrero Fernandez, F.A. Guerrero Pena, T. Ren, A. Cunha, FERAtt: Facial expression recognition with attention net, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, 2019.

[20] Y. Jiao, Y. Niu, Y. Zhang, F. Li, C. Zou, G. Shi, Facial attention based convolutional neural network for 2d+ 3d facial expression recognition, in: 2019 IEEE Visual Communications and Image Processing (VCIP), IEEE, 2019, pp. 1–4.

[21] S. Woo, J. Park, J.-Y. Lee, I. So Kweon, Cbam: Convolutional block attention module, in: Proceedings of the European Conference on Computer Vision (ECCV), 2018, pp. 3–19.

[22] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A.N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017, pp. 5998–6008.

[23] H. Ling, J. Wu, J. Huang, J. Chen, P. Li, Attention-based convolutional neural network for deep face recognition, Multimedia Tools Appl. 79 (9) (2020) 5595–5616.

[24] G. Wang, W. Wang, J. Wang, Y. Bu, Better deep visual attention with reinforcement learning in action recognition, in: 2017 IEEE International Symposium on Circuits and Systems (ISCAS), IEEE, 2017, pp. 1–4.

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

[26] J. Li, C. Soladie, R. Seguier, Ltp-ml: micro-expression detection by recognition of local temporal pattern of facial movements, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 634–641.

[27] A. Davison, W. Merghani, C. Lansley, C.-C. Ng, M.H. Yap, Objective micro-facial movement detection using facs-based regions and baseline evaluation, in: 2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018), IEEE, 2018, pp. 642–649.

[28] H. Lu, K. Kpalma, J. Ronsin, Micro-expression detection using integral projections, in: WSCG 2017, Václav Skala-UNION Agency, 2017.

[29] A.K. Davison, C. Lansley, N. Costen, K. Tan, M.H. Yap, Samm: A spontaneous micro-facial movement dataset, IEEE Trans. Affect. Comput. 9 (1) (2016) 116–129.

[30] A. Davison, W. Merghani, M. Yap, Objective classes for micro-facial expression recognition, J. Imaging 4 (10) (2018) 119.

[31] X. Huang, G. Zhao, X. Hong, W. Zheng, M. Pietikäinen, Spontaneous facial micro-expression analysis using spatiotemporal completed local quantized patterns, Neurocomputing 175 (2016) 564–578.

[32] X. Li, J. Yu, S. Zhan, Spontaneous facial micro-expression detection based on deep learning, in: 2016 IEEE 13th International Conference on Signal Processing (ICSP), IEEE, 2016, pp. 1130–1134.

[33] Z. Zhang, T. Chen, H. Meng, G. Liu, X. Fu, Smeconvnet: A convolutional neural network for spotting spontaneous facial micro-expression from long videos, IEEE Access 6 (2018) 71143–71151.

[34] P.Y. Simard, D. Steinkraus, J.C. Platt, et al., Best practices for convolutional neural networks applied to visual document analysis, in: Icdar, Vol. 3, 2003.

[35] S. Thuseethan, S. Rajasegarar, J. Yearwood, Detecting micro-expression intensity changes from videos based on hybrid deep CNN, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, Springer, 2019, pp. 387–399.

[36] K. Zuiderveld, Contrast limited adaptive histogram equalization, in: Graphics Gems IV, Academic Press Professional, Inc., USA, 1994, pp. 474–485.

[37] M. Yang, L. Van Gool, L. Zhang, Sparse variation dictionary learning for face recognition with a single training sample per person, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 689–696.

[38] M. Yang, D. Dai, L. Shen, L. Van Gool, Latent dictionary learning for sparse representation based classification, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 4138–4145.

[39] Y. Peng, D. Meng, Z. Xu, C. Gao, Y. Yang, B. Zhang, Decomposable nonlocal tensor dictionary learning for multispectral image denoising, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2014, pp. 2949–2956.

[40] C. Lu, J. Shi, J. Jia, Online robust dictionary learning, in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 415–422.

[41] J. Xing, J. Gao, B. Li, W. Hu, S. Yan, Robust object tracking with online multi-lifespan dictionary learning, in: Proceedings of the IEEE International Conference on Computer Vision, 2013, pp. 665–672.

[42] W.-J. Yan, Q. Wu, Y.-J. Liu, S.-J. Wang, X. Fu, Casme database: a dataset of spontaneous micro-expressions collected from neutralized faces, in: 2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG), IEEE, 2013, pp. 1–7.

[43] W.-J. Yan, X. Li, S.-J. Wang, G. Zhao, Y.-J. Liu, Y.-H. Chen, X. Fu, Casme ii: An improved spontaneous micro-expression database and the baseline evaluation, PLoS One 9 (1) (2014) e86041.

[44] F. Qu, S.-J. Wang, W.-J. Yan, H. Li, S. Wu, X. Fu, Cas(me)^2: A database for spontaneous macro-expression and micro-expression spotting and recognition, IEEE Trans. Affect. Comput. 9 (4) (2017) 424–436.

[45] D. Borza, R. Itu, R. Danescu, Real-time micro-expression detection from high speed cameras, in: 2017 13th IEEE International Conference on Intelligent Computer Communication and Processing (ICCP), IEEE, 2017, pp. 357–361.

[46] X. Li, X. Hong, A. Moilanen, X. Huang, T. Pfister, G. Zhao, M. Pietikäinen, Towards reading hidden emotions: A comparative study of spontaneous micro-expression spotting and recognition methods, IEEE Trans. Affect. Comput. 9 (4) (2017) 563–577.

[47] J. Ding, Z. Tian, X. Lyu, Q. Wang, B. Zou, H. Xie, Real-time micro-expression detection in unlabeled long videos using optical flow and lstm neural network, in: International Conference on Computer Analysis of Images and Patterns, Springer, 2019, pp. 622–634.

[48] J. Li, C. Soladie, R. Seguier, S.-J. Wang, M.H. Yap, Spotting micro-expressions on long videos sequences, in: 2019 14th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2019), IEEE, 2019, pp. 1–5.

[49] S.-T. Liong, K. Wong, Micro-expression recognition using apex frame with phase information, in: 2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), IEEE, 2017, pp. 534–537.

Madhumita Abhijeet Takalkar is a Ph.D. candidate at the University of Technology Sydney, Australia. She holds a M.E. degree in Computer Science and Engineering from Government Engineering College, Aurangabad (MS), India and B.E. degree in Information Technology from P.E.S. College of Engineering, Aurangabad (MS), India (Affiliated to Dr. Babasaheb Ambedkar Marathwada University, Aurangabad (MS)). She has particular research interests in deep learning, digital image processing, computer vision and multimedia applications.

Selvarajah Thuseethan received the B.Sc. degree from University of Jaffna, Sri Lanka, and currently pursuing Ph.D. degree from the School of Information Technology, Deakin University, Geelong, Australia. He has been severed as a probationary lecturer with Sabaragamuwa University of Sri Lanka. His current research interests include emotion recognition, computer vision, machine learning, and pattern recognition.

Sutharshan Rajasegarar received the Ph.D. degree from The University of Melbourne, Melbourne, VIC, Australia. He was a Research Fellow with the Department of Electrical and Electronic Engineering, The University of Melbourne, and a Researcher in machine learning with the National ICT Australia. He is currently a Senior lecturer with the School of Information Technology, Deakin University, Geelong, Australia. His current research interests include anomaly/outlier detection, distributed machine learning, spatio-temporal estimations, pattern recognition, computer vision, health analytics, and wireless communication.

Zenon Chaczko received the bachelor's degree (Hons.) in Computer Science and the Ph.D. degree in Computer and Software Engineering. He is currently a Senior Lecturer and the Director of the ICT Programs with the FEIT, University of Technology Sydney (UTS). He is also a Core Member of GBDTC-Global Big Data Technologies Center. Before UTS, he worked full-time for over 20 years in the software and system engineering industry. He held various visiting positions: visiting professorships at the University of Arizona, Tucson, USA, IPN, Mexico City, Mexico, the University of Las Palmas, Canary Islands, Spain, the Wroclaw University of Technology, Poland, and the University of Applied Science and Engineering, Hagenberg/Linz, Austria. He was a keynote speaker at multiple conferences. He has been actively conducting research in his fields and has published over 250 research articles. His research interests include artificial intelligence, software systems engineering, cloud and ambient computing, augmented reality, and complex software systems.

Min Xu is currently an Associate Professor at University of Technology, Sydney. She has received Ph.D. degree from University of Newcastle, Australia, M.S. degree from National University of Singapore and B.E. degree from University of Science and Technology of China. Her research interests include multimedia data (video, audio and text) analytics and computer vision. She has published over 100 research papers in high quality international journals and conferences. Dr. Xu also serves several journals and conferences as reviewer including IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, ACM Multimedia conference. She is an Associate Editor of Journal of Neurocomputing.

John Yearwood received the B.Sc. degree from Monash University, Australia, the M.Sc. degree from Sydney University, Australia, and the Ph.D. degree from RMIT University, Australia. In 1989, he was a Lecturer with the School of Information Technology and Mathematical Science, University of Ballarat, Australia, where he was appointed as a Professor in 2007. He is currently a Professor and the Head with the School of Information Technology, Deakin University, Australia. He has authored over 140 refereed journals, book chapters, and conference articles. His research interest includes modern optimization theory and techniques and their applications in pattern recognition, signal processing, and decision support systems.