IET Image Processing

Research Article

IET Journals
The Institution of Engineering and Technology

Regional adaptive affinitive patterns (RADAP) with logical operators for facial expression recognition

ISSN 1751-9659
Received on 16th June 2018
Revised 5th January 2019
Accepted on 11th February 2019
E-First on 18th March 2019
doi: 10.1049/iet-ipr.2018.5683
www.ietdl.org

Murari Mandal[1], Monu Verma[1], Sonakshi Mathur[1], Santosh Kumar Vipparthi[1] ✉, Subrahmanyam Murala[2], Deveerasetty Kranthi Kumar[3]

[1]*Department of Computer Science and Engineering, Malaviya National Institute of Technology, Jaipur, India*
[2]*Department of Electrical Engineering, Indian Institute of Technology Ropar, Roopnagar, India*
[3]*Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, People's Republic of China*
✉ *E-mail: skvipparthi@mnit.ac.in*

Abstract: Automated facial expression recognition plays a significant role in the study of human behaviour analysis. In this study, the authors propose a robust feature descriptor named regional adaptive affinitive patterns (RADAP) for facial expression recognition. The RADAP computes positional adaptive thresholds in the local neighbourhood and encodes multi-distance magnitude features which are robust to intra-class variations and irregular illumination variation in an image. Furthermore, they established cross-distance co-occurrence relations in RADAP by using logical operators. They proposed XRADAP, ARADAP, and DRADAP using xor, adder and decoder, respectively. The XRADAP engrains the quality of robustness to intra-class variations in RADAP features using pairwise co-occurrence. Similarly, ARADAP and DRADAP extract more stable and illumination invariant features and capture the minute expression features which are usually missed by regular descriptors. The performance of the proposed methods is evaluated by conducting experiments on nine benchmark datasets Cohn–Kanade+ (CK+), Japanese female facial expression (JAFFE), Multimedia Understanding Group (MUG), MMI, OULU-CASIA, Indian spontaneous expression database, DISFA, AFEW and Combined (CK+, JAFFE, MUG, MMI & GEMEP-FERA) database in both person dependent and person independent setup. The experimental results demonstrate the effectiveness of the proposed method over state-of-the-art approaches.

1 Introduction

Humans communicate their intentions and emotions through external agents such as speech, gestures, and facial expressions. Facial appearances provide valuable information about the affective state, cognitive activity, personality, and psychology of a person [1]. Automatic facial expression recognition (FER) presents a non-intrusive approach to analyse human affective behaviour. The FER system plays an important role in human–computer interaction, surveillance, deceit or lie detection, behavioural profiling, fatigue monitoring, data-driven animation, and health-care applications. Due to such a wide range and continuous evolution in these applications, automated emotion analysis has attracted increasing attention in the past few decades.

Ekman and Friesen [2] identified different facial expressions by developing the facial action coding system (FACS), which was further simplified to emotional FACS [3]. These studies established that emotions of people across different cultures are characterised by similar facial expressions. Also, a set of six prototype emotions named anger, happy, sad, surprise, fear, and disgust were identified, which can be recognised based on the facial expressions. However, there are many factors that create challenges in developing an automated FER system such as variation in facial expressions, illumination changes, age, gender, ethnicity variation etc.

A general FER system takes the flow of a classical pattern recognition problem, which includes the task of image acquisition, developing pre-processing techniques, robust feature extraction, classification, and post-processing techniques. The effectiveness of such a framework is largely dependent on the accurate feature extraction and classification technique. Inadequate feature extraction would degrade the performance even after using the best classification techniques. Therefore, designing an appropriate feature descriptor is essential for a robust FER system.

Feature extraction techniques can be broadly divided into two categories: hand-crafted features and learned features [4]. The hand-crafted features are predesigned to extract relevant facial expressions whereas the learned features are encoded by using deep neural networks. The deep learning approaches [5–17] collectively learn the appropriate features and the classification weights to recognise the facial expression. These features are either extracted from an entire facial region, i.e. global features or from a specific region of interest, i.e. local features. The local features are usually selected corresponding to the action units (AUs). Furthermore, feature extraction can also be divided into spatial and spatiotemporal representation from single image and frame sequences, respectively [18]. The hand-crafted features in the literature can be divided into geometric and appearance-based features. The geometric features [19, 20] represent the face by encoding the shape, deformation, corner, distances, contour, and other geometric properties. These features are calculated on the basis of related fiducial point's locations. The geometric features fail to identify the minute characteristics such as ridges and skin texture changes and are dependent on reliable and accurate feature detection and tracking. In addition, pre-processing techniques are required to localise various facial components before the extraction of facial features.

The appearance features detect a more complete set of features and are resistant to noise. Furthermore, the facial image can be divided into non-overlapping regions to extract features, which increase the performance of the system [21]. The selection of salient grids, size, and location of the grids directly affects the recognition accuracy of the FER system. The local feature descriptors [21–25] exploit the gradient variations to detect the facial expressions in an image. Similarly, the edge-based feature descriptors [26–29] use the filter masks uniformly over an image and select the salient magnitude responses to detect the directional patterns.

In our opinion, the effectiveness of the existing feature descriptors used for FER applications is impeded by the following shortcomings:

i. The edge-based features [local directional ternary pattern (LDTerP) [25], local directional number (LDN) [26], and local directional texture patterns (LDTP) [29]) generate unstable patterns in smooth regions.
ii. The irregular illumination changes in the facial images affect the feature representation capability of the local binary patterns (LBP) based local descriptors.
iii. There is always a trade-off between extracting detailed features and being robust to intra-class variations. Most of the descriptors lack the inbuilt capability to extract the salient features and being robust to noise (irrelevant features) at the same time.
iv. The CNNs sometime tend to learn the abstract facial features rather than the specific finer changes in the face image, which characterise a particular expression.

Aiming to address the aforementioned shortcomings, in this study, we propose a novel feature descriptor, regional adaptive affinitive patterns (RADAP), for FER. The main contribution of the proposed work is as follows:

i. We developed an adaptive global threshold generation technique for encoding RADAP at multiple distances within the local neighbourhood. It captures both the local and globally invariant features in the neighbourhood.
ii. Since the RADAP assimilates the global changes in the local neighbourhood, thus, the responses are better encoded to become robust to noise and uneven illumination changes.
iii. Traditional unsupervised threshold-based methods [30] compute the threshold based on the individual pixel value. A pixel's intensity value does not usually represent the normalised intensity within a region. This may lead to the inclusion of outliers during the thresholding process. Furthermore, the illumination variations in an image cannot be adequately represented using a single global threshold. To address these problems, RADAP determines the thresholds adaptively by identifying the relationship between the pixels in the local region.
iv. It also incorporates the multi-distance information in the local neighbourhood to detect finer changes.
v. The proposed method is further extended by establishing cross-distance co-occurrence relationship using logical operators. We used the xor, adder and decoder operators over the RADAP patterns to propose the XRADAP, ARADAP and DRADAP descriptors, respectively. To the authors' knowledge, such cross-distance correlation information has not been captured in this manner in the literature.

The feature vectors are used with multiclass support vector machines (SVM) for expression classification. The overall procedure for FER using the RADAP descriptors is shown in Fig. 1. The robustness of the proposed methods is validated by conducting experiments over nine standard facial expression datasets.

The rest of the paper is organised as follows: In Section 2, the related state-of-the-art methods are discussed. In Section 3, the proposed descriptor is discussed and an algorithmic representation of the proposed descriptor is given. We discuss the classification technique along with performance measures in Section 4. In Section 5, we delineate the experimental results and discuss the outcome along with state-of-the-art techniques. Finally, based on the above work, we conclude and provide future work with some suggestions in Section 6.

2 Related work

Zhang *et al.* [31] extracted 34 fiducial points from the face image as landmark points. These fiducial points are used to extract the geometric features by modelling the salient facial locations and shape information. Valstar *et al.* [32] proposed to track the facial points and detect the AUs in the face image. The facial expressions can be recognised based on the detected AUs in the image.

Appearance-based methods have been widely used to measure the physical appearance of a facial image. Especially, the local feature descriptor-based methods for facial appearance analysis have gained popularity due to their ease of implementation, pose invariance, and robustness to illumination. These methods capture the spatial topology of the image in the local neighbourhood. Shan *et al.* [21] used LBP to extract the facial features for expression recognition. Lai *et al.* [22] employed a two-stage feature extraction; first, they retrieved threshold LBP responses and then applied centre symmetric-LBP. In local directional patterns (LDP) [28], the local edge responses in eight directions were computed using eight different masks. Extraction of the salient directional responses increases the discriminative capability of the descriptors. Rivera *et al.* [26] represented the directional information by encoding a more discriminative LDN patterns. They also [29] proposed to extract the texture information by identifying the principal directions and encoded the intensity variation of the principal directions into response numbers called LDTP. Ryu *et al.* [27] used the ternary patterns to extract the directional information and designed a multi-level grid-based approach to characterise the coarse and finer features separately. The coarse grids are used for stable codes, which are closely related to non-expressions, whereas the finer grids are used for active codes which are closely related to expressions.

Lee *et al.* [33] proposed to use the sparse representation of images to reduce the intra-class variations of expressions. Mohammadzade and Hatzinakos [34] introduced the concept of expression subspace which represents a particular expression with one subspace and new expressions can be synthesised from an image by applying projections into different expression subspaces. Hybrid methods incorporate various techniques from geometric and appearance-based approaches to attain enhanced performance.

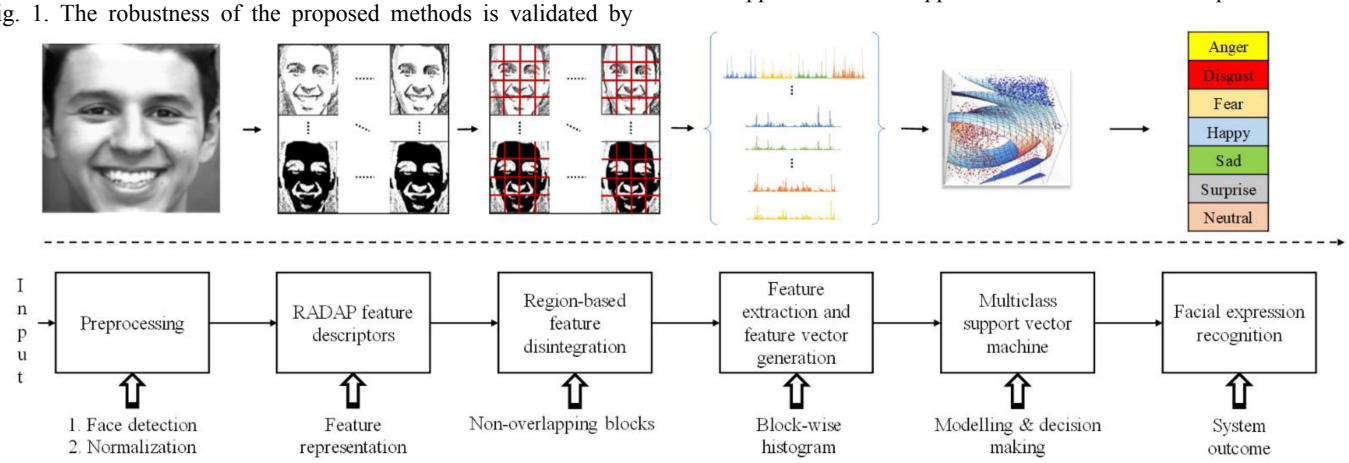

Fig. 1 *Overview of the proposed method. For each image, first, the face region is detected and normalised. Then, we calculate the RADAP features and extract the feature vector by concatenating the block-wise histogram of the response codes. The feature vector is used to model a multiclass SVM to classify the facial*

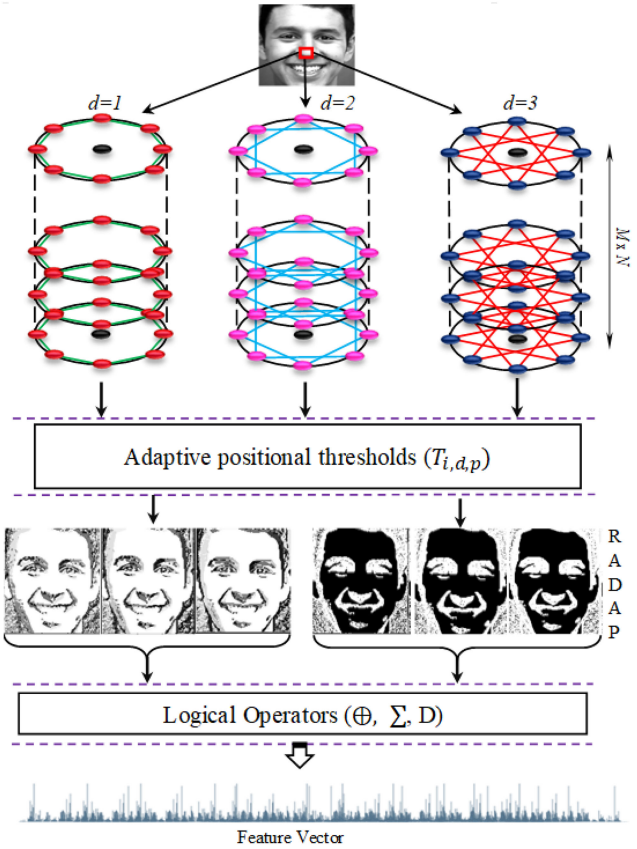

Fig. 2 *Overview of the proposed descriptors. The positional adaptive thresholds are computed at a distance*
(a) $d=1$, ***(b)*** $d=2$, ***(c)*** $d=3$. At each distance, eight thresholds ($T_{i,d,p}$) are generated where $i \in [0, p-1]$

Zhang *et al.* [35] proposed to capture facial movement features based on distance features. These distance features are computed by extracting the salient patch-based Gabor features. Happy *et al.* [36] extracted the salient patches from various active facial patches during the emotion elicitation. They further extracted LBP features from the salient patches to generate the feature vector and classified the expression images using the SVM classifier. Furthermore, many significant works for the advancement of FER systems were done in [37–39].

More recent works in facial expression analysis have used deep learning approaches to solve the recognition problem. The deep learning methods learn both the feature extraction and the network weight parameters for accurate classification using the training data. Burkert *et al.* [5] designed a deep convolutional neural network (CNN) inspired by GoogleNet and introduced a parallel feature extraction (FeatEx) block to extract features at different scales. Mollahosseini *et al.* [6] applied the concept of network in network architecture and added inception layers after two traditional CNN layers. The inception layers are concatenated as output and connected to the fully connected layers. Barsoum *et al.* [7] trained a customised VGG13 network to verify different crowd-sourced label distribution techniques and undertook facial expression classification as a case study. Hasani and Mahoor [8] designed a network consisting of 3D inception-ResNet layers followed by an Long short-term memory (LSTM) module. These layers extract both the spatial and temporal relations in the face image and frame sequences in the video, respectively. A two-step training method was proposed by Ding *et al.* [9] where, in the first stage, the convolutional layer weights are regularised and in the second stage, the fully connected layers are added to the pre-tuned convolution layers and train the complete network to learn the optimal classification parameters. A combination of multiple CNN architectures is yielding better classification accuracy. With this consideration, Pons *et al.* [10] proposed to improve the FER accuracy by supervised learning of committee of CNNs. Kim *et al.* [11] combined the decisions from a hierarchical committee of CNNs and hand-crafted hierarchical decision rule. Other CNN-based systems such as VGG [12], ResNet [13], DTAGN [14], DTAGN-Joint [15], spatio-temporal [16] and GCNet [17] have also achieved accelerated growth in the field of FER.

3 Proposed methods

In this study, we proposed a new feature descriptor RADAP for FER. The local edge gradients have been proven very effective in characterising the facial appearances of different expressions. In this study, we propose to capture both the local changes and the globally invariant features in the neighbourhood. The properties of the RADAP are as follows: (i) the positional adaptive threshold strengthens its ability to become invariant to un-even illumination changes and also generate more stable patterns in smooth regions, (ii) it encodes the multi-distance sign and magnitude patterns to improve the discriminative ability in the local facial region, (iii) the pairwise co-occurrence and fused cross-distance patterns (XRADAP, ARADAP and DRADAP) enhance the proposed descriptors ability to extract the salient expression features and being robust to noise (irrelevant features) at the same time.

To compute the RADAP, we extracted the adaptive positional thresholds for a facial image. These positions are indexed by considering multiple distances among the neighbours for a reference pixel. These thresholds are calculated by considering the global mean of all the magnitude responses at that position. Thus, it adaptively adjusts the threshold parameters in the local neighbourhood for different images. This is in contrast to the existing methods where the global patterns do not influence the features captured at the local regions. For each distance d, we extract eight positional thresholds and thus, 24 positional thresholds are computed for $d \in [1, D]$, where $D = 3$. Let $I(a, b)$ be a greyscale image of size $M \times N$, where $a \in [1, M]$ and $b \in [1, N]$. If at each location in the image, the p neighbourhood pixels are situated at a radius r. The positional adaptive threshold $T_{i,d,p}$ at distance d and position i is calculated using (1). The same is shown in Fig. 2

$$T_{i,d,p} = \frac{1}{(M-2)\times(N-2)} \sum_{j=1}^{M-2} \sum_{k=1}^{N-2} \left| G_{\beta_{i,d,p}}(j,k) - G_i(j,k) \right|, \quad (1)$$

where $i \in [0, p-1]$, $G_\Theta = I(R_1{:}R_2, C_1{:}C_2)$ and R_1, R_2, C_1, C_2 are computed using (2)–(5)

$$R_1 = \begin{cases} 1, & \text{if } \Theta \in \{ \bmod(\alpha, p)\}_{\alpha=1}^{D}, \\ 3, & \text{if } \Theta \in \{ \bmod(\alpha+4, p)\}_{\alpha=1}^{D}, \\ 2, & \text{otherwise}, \end{cases} \quad (2)$$

$$C_1 = \begin{cases} 1, & \text{if } \Theta \in \{ \bmod(\alpha+2, p)\}_{\alpha=1}^{D}, \\ 3, & \text{if } \Theta \in \{ \bmod(\alpha+6, p)\}_{\alpha=1}^{D}, \\ 2, & \text{otherwise}, \end{cases} \quad (3)$$

$$R_2 = \begin{cases} M & \text{if } \Theta \in \{ \bmod(\alpha+4, p)\}_{\alpha=1}^{D}, \\ M-2 & \text{if } \Theta \in \{ \bmod(\alpha, p)\}_{\alpha=1}^{D}, \\ M-1 & \text{otherwise}, \end{cases} \quad (4)$$

$$C_2 = \begin{cases} N, & \text{if } \Theta \in \{ \bmod(\alpha+6, p)\}_{\alpha=1}^{D}, \\ N-2, & \text{if } \Theta \in \{ \bmod(\alpha+2, p)\}_{\alpha=1}^{D}, \\ N-1, & \text{otherwise}. \end{cases} \quad (5)$$

The RADAP is computed by encoding the *mRADAP* and *sRADAP* using $T_{i,d,p}$ as given in (6)–(10)

$$\text{mRADAP}_{p,d}(a,b) = \sum_{i=0}^{p-1} \psi_1(x_{\beta_{i,d,p}} - x_i) \times 2^i, \quad (6)$$

IET Image Process., 2019, Vol. 13 Iss. 5, pp. 850-861

$$\text{sRADAP}_{p,d}(a,b) = \sum_{i=0}^{p-1} \psi_2(x_{\beta_{i,d,p}} - x_i) \times 2^i, \tag{7}$$

$$\psi_1(x) = \begin{cases} 1, & \text{if } |x| \geq T_{i,d,p}, \\ 0, & \text{otherwise}, \end{cases} \tag{8}$$

$$\psi_2(x) = \begin{cases} 1, & \text{if } x \geq 0, \\ 0, & \text{otherwise}, \end{cases} \tag{9}$$

$$\beta_{i,d,p} = \begin{cases} \beta^1_{i,d,p}, & \text{if } \beta^1_{i,d,p} - 1 < p - 1, \\ \beta^2_{i,d,p}, & \text{otherwise}, \end{cases} \tag{10}$$

where $\beta^1_{i,d,p} = i + d$ and $\beta^2_{i,d,p} = i + d - p$. The sample calculation of the RADAP codes is shown in Fig. 3.

3.1 XRADAP

Most of the existing descriptors lack the inbuilt capability to extract the salient features and being robust to intra-class variations at the same time. The XRADAP is encoded to extract the pairwise co-occurrence patterns to engrain this quality in the RADAP descriptor. These patterns are generated by applying xor-based transformation over RADAP. The response map of the XRADAP over a given sRADAP bit-pattern is shown in Table 1. From Table 1, it is clear that the proposed transformation identifies the co-occurrence between patterns extracted at different distances in the local neighbourhood. This operation enables discrimination between similar and dissimilar cross-distance RADAP which improves the saliency of the RADAP features and results in $2 \times ({}^{D}C_2)$ feature maps, which is equal to the feature dimension of the RADAP. The XRADAP features for sample facial expressions are shown in Fig. 4. The XRADAP is computed using (11)

$$\text{XRADAP}_{\ell,p,d}(a,b) = \sum_{i=0}^{p-1} \left\{ \psi_\ell(x_{\beta_{i,d,p}} - g_i) \oplus \psi_\ell(x_{\beta_{i,\eta,p}} - g_i) \right\} \times 2^i, \tag{11}$$

where $\eta = \mod(d,3) + 1$ and $\ell \in \{1,2\}$.

3.2 ARADAP

We also extracted cross-distance fused patterns (ARADAP) over the RADAP using adder-based transformation. The idea is to extract more stable and illumination invariant features from the already robust RADAP patterns. It can be seen from Fig. 4 that the ARADAP substantially removes the noise (irrelevant features) in the RADAP and further enhances the salient expression changes even in different illumination conditions. The computation of an adder-based response is shown in Table 1. At each neighbourhood position, the adder response is generated by summing up the hamming weights from multiple distances. Furthermore, an impulse function is applied over these responses to generate the ARADAP patterns. The fused cross-co-occurrence features are extracted to distinguish maximum and minimum variation in the local facial region. Therefore, the strongest response is produced when all three features have a hamming weight of 'one'. Hence, ARADAP is more robust to handle the noise present in a local region. It generates $2 \times (D+1)$ feature maps. The mathematical representation of ARADAP is given in (12) and (13)

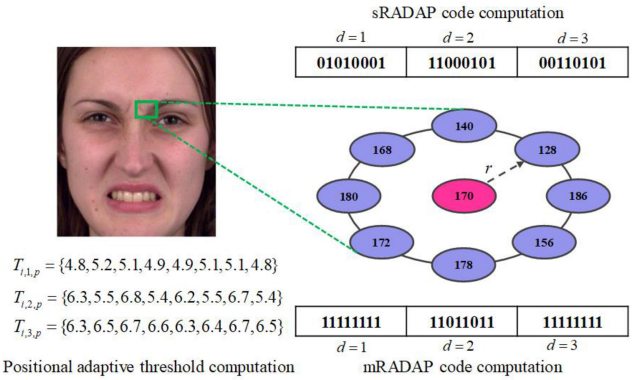

Fig. 3 *RADAP code computation. We calculate positional adaptive threshold from the whole image and compute the mRADAP using those thresholds. We show an example of* 3×3 *image patch and compute its sRADAP and mRADAP patterns*

$$\text{ARADAP}_{\ell,z_1}(a,b) = \sum_{i=0}^{p-1} \text{AR}_{\ell,z_1,i}(a,b) \times 2^i, \tag{12}$$

$$AR_{\ell,z_1,i}(a,b) = \begin{cases} 1, & \text{if } \sum_{d=1}^{D} \psi_\ell(x_{\beta_{i,d,p}} - x_i) \Big|_{i \in [0,p-1]} = (z_1 - 1), \\ 0, & \text{otherwise}, \end{cases} \tag{13}$$

where $z_1 \in [1, D+1]$.

3.3 DRADAP

Similar to ARADAP, we used the decoder-based transformation over RADAP which captures the minute expression features usually missed by regular descriptors. However, the DRADAP still remains invariant to uneven illumination changes in the facial image and the same is shown in Fig. 4. It is calculated by selecting the 3-bit stream (given weighted precedence) at each plane from multiple distance responses as given in Table 1. The DRADAP gives weighted precedence to the distance feature responses where $d=1$ and $d=3$ have the highest and lowest precedence, respectively. Finally, an impulse function is utilised to extract the DRADAP patterns. This approach maintains a balance between robustness to noise and finer detection of co-occurrence relationships in the local neighbourhood. The proposed method in total generates $2 \times (2^D)$ feature maps. The DRADAP is computed using (14) and (15)

$$\text{DRADAP}_{\ell,z_2}(a,b) = \sum_{i=0}^{p-1} DR_{\ell,z_2,i}(a,b) \times 2^i, \tag{14}$$

$$\text{DR}_{\ell,z_2,i}(a,b) = \begin{cases} 1, & \text{if } \sum_{d=1}^{D} \psi_\ell(x_{\beta_{i,d,p}} - x_i) \times 2^{D-d} \Big|_{i \in [0,p-1]} = (z_2 - 1), \\ 0, & \text{otherwise}, \end{cases} \tag{15}$$

where $z_2 \in [1, 2^D]$.

Table 1 Xor, adder & decoder response map for a sample multi-distance sRADAP bit pattern

$d=1$	$d=2$	$d=3$	Xor	Adder	Decoder
0	1	0	1 1 0	1	2
1	1	0	0 1 1	2	6
0	0	1	0 1 1	1	1
1	0	1	1 1 0	2	5
0	0	0	0 0 0	0	0
0	1	1	1 0 1	2	3
0	0	0	0 0 0	0	0
1	1	1	0 0 0	3	7

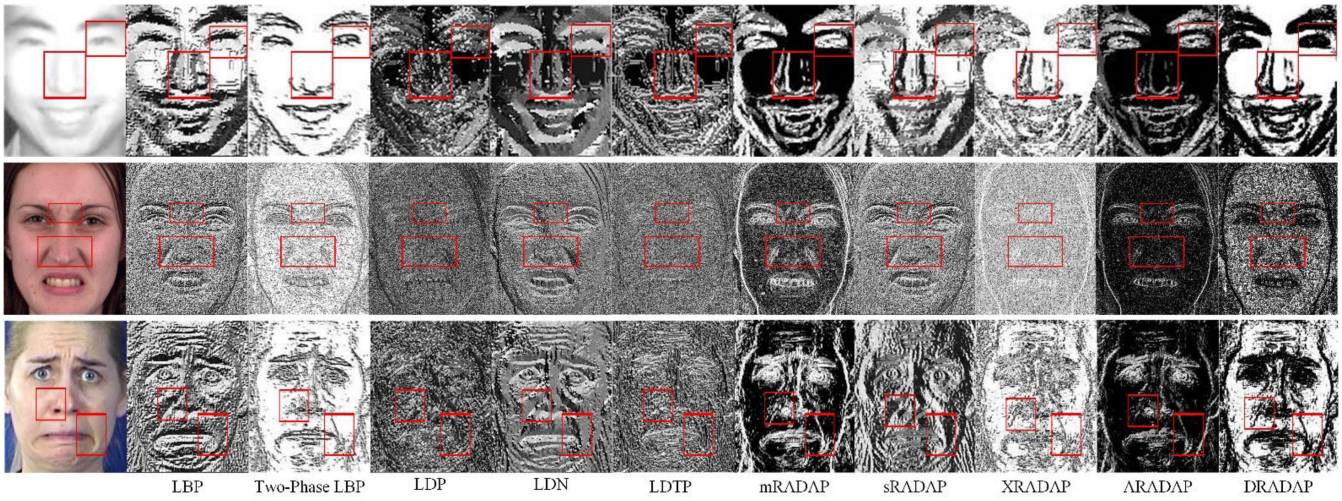

Fig. 4 *Input image and its response maps generated by the existing state-of-the-art descriptors and the proposed methods RADAP, XRADAP, ARADAP, and DRADAP, respectively. The highlighted regions show the qualitative difference between the RADAP features and existing feature descriptors. For the first input image, the proposed methods (RADAP, XRADAP, ARADAP, and DRADAP) are able to extract more refined salient features (in the mouth, eyes, and nose regions) as compared to the features detected by LDTerP, LBP, LDN, and LDTexP. Moreover, in the second, third, and fourth image, the proposed RADAP is more robustly preserving the features in the expressive regions as compared to the other descriptors*

3.4 Feature vector

In this study, the resultant feature response maps are divided into equal-sized non-overlapping blocks. As stated in [24], the block-level feature extraction provides more discriminative information for facial images. The block level approach helps to identify variation in different levels of locality, i.e. pixel level and region level. Therefore, we have generated the final feature vector by concatenating the block level vector responses. If the descriptor response map of size $M \times N$ is divided into S non-overlapping regions of equal size $\{R_1, R_2, R_3, \ldots, R_S\}$ and the size of each block is $B \times B$. Then, the number of blocks in a row is $R_{\text{blk}} = \lceil M/B \rceil$ and the number of blocks in a column is $C_{\text{blk}} = \lceil N/B \rceil$. The feature vector $\boldsymbol{H}_{\text{FER}}$ for facial expression is computed using (16)

$$\boldsymbol{H}_{\text{FER}} = \left\{ \{H(F_m, W, v)\}_{m=1}^{\Gamma} \right\}_{v=1}^{S}, \tag{16}$$

where Γ is the total number of response maps generated from a feature descriptor, F_m is the transformed response map, $W \in [0, 255]$ and $H(\cdot, \cdot, \cdot)$ is computed using (17)–(19)

$$H(F_m, W, v) = \sum_{a=\Omega_1}^{\Omega_1+B} \sum_{b=\Omega_2}^{\Omega_2+B} \delta(F_m(a, b) - W), \tag{17}$$

$$\Omega_1 = 1 + (\lfloor v/R_{\text{blk}} \rfloor \times B), \tag{18}$$

$$\Omega_2 = (\bmod (v-1, R_{\text{blk}}) \times B) + 1, \tag{19}$$

where Ω_1 and Ω_2 give the row and column index for each block.

3.5 Comparison with existing feature descriptors

Directional edge response-based methods [26–29] have shown good performance in extracting relevant facial features for expression classification. In LDP [28], LDN [26], LDTP [29] and LDTerP [27], the authors have used eight predefined compass masks (Robinson & Kirsch) to calculate the edge responses. They generated their coded patterns by analysing the salient directional index positions in the local neighbourhood. However, such approaches are also very susceptible to uneven illumination changes in the facial region and fail to extract accurate features in smooth regions, which can be observed in Fig. 4. The RADAP addresses these problems by determining the thresholds adaptively based on the relationship between the pixels in the local region. It assimilates the global changes in the local neighbourhood, thus, the responses are better encoded to become robust to intra-class variations and uneven illumination changes. Furthermore, the co-occurrence among the RADAP features at multiple distances is computed to extract more stable features (invariant to factors such as irrelevant features, variations in the location of active regions for the same expression), which can enable more accurate recognition accuracy in the FER system. From Fig. 4, we can see that these logical operator-based relations significantly improve the feature saliency maps of the RADAP descriptors.

The CNN-based approaches for FER learn the expression features from the facial images and perform classification based on the learned features. Although the deep learning techniques have achieved noticeable success in image classification tasks, their effectiveness in FER is sometimes impeded by the fact that the CNNs tend to learn the abstract facial features rather than the specific finer changes in the face image, which characterise a particular expression. Also, in its present form, it is very difficult to interpret the CNN models proposed to solve the problem of FER. Therefore, the proposed designed features such as RADAP are still relevant for automatic recognition of emotions.

4 Classification and performance measure

We perform classification by training an SVM model using the feature vector as computed in the previous section. SVM attempts to identify the optimal hyperplane between two classes to maximise the inter-class margin [40]. Let $\{(A_\kappa, B_\kappa), \kappa = 1, \ldots, w\}$ be the set of labelled training samples for $A_\kappa \in \boldsymbol{C}^n$ and $B_\kappa \in \{+1, -1\}$, where $\boldsymbol{C}^n$ is the feature vector for each class. A new test sample ϖ can be classified using (20)

$$f(\varpi) = \text{sign}\left(\sum_{\kappa=1}^{w} \phi_\kappa B_\kappa Z(A_\kappa, \varpi) + \rho \right), \tag{20}$$

where $Z(\cdot, \cdot)$ is the kernel function, ϕ_κ denotes the Lagrange's multipliers for a dual optimisation problem and the bias parameter is denoted by ρ. The multi-class classification model is designed by employing $\wp(\wp - 1)/2$ binary SVM models using the one-versus-one coding design, where $\wp$ is the total number of class labels. In this study, the SVM models were implemented using linear kernel function and the performance was evaluated using k-fold cross validation.

The performance of the FER system is calculated in terms of recognition accuracy, which is computed by using (21)

$$\text{Recognition accuracy} = \frac{\text{total no. of correctly classified samples}}{\text{total no. of samples}}. \quad (21)$$

5 Experimental results and discussions

We validate the effectiveness of the proposed method by conducting nine experiments on Cohn–Kanade+ (CK+) [41, 42], Japanese female facial expression (JAFFE) [43], Multimedia Understanding Group (MUG) [44], MMI [45, 46], OULU-CASIA [47], Indian spontaneous expression database (ISED) [48], DISFA [49], AFEW [50, 51] and the combined (CK+, JAFFE, MUG, MMI, and GEMEP-FERA) database for FER. In our experiments, the proposed descriptors are applied over pre-processed (facial detection & size normalisation to 120×20) image responses. The generated feature maps are apportioned into blocks of size $B \times B$, where $B = 8$. Contrary to the approaches in the literature [27, 29], the pre-processing steps in our experiment do not involve any manual cropping of facial regions. We use the standard Viola Jones [52] face detection algorithm to get the facial region. This approach helps us mirror the real-life application scenarios where the facial expressions are classified based on the detected facial region. Furthermore, instead of using variable block sizes to achieve optimal performance in different facial expression datasets, we have used a uniform block size for all the datasets. This alleviates the problem of parameter selection and non-universality of the FER system. It also ensures a fair comparison between the proposed descriptors with the existing feature descriptors. The facial images were arranged by selecting the most expressive frames from each video sequence in the dataset. This setup has been widely adopted in the literature [26, 27, 29].

In this study, we have performed the experiments on both person dependent (PD) and person independent (PI) cross-validation schemes to evaluate the effectiveness of the proposed descriptor.

PD scheme: In PD cross validation, image sets are randomly divided into N parts.

The $N-1$ parts are used as a training set and rest is used as a testing set. We have randomly divided the dataset into a ratio of 80:20 and selected the training and testing set images, respectively. The final recognition accuracy is calculated by taking the average of the accuracy produced after five iterations.

PI scheme: In the PI scheme, it is ensured that a person's facial expressions are not divided into training and testing sets. All the expressions of a single or a set of subjects are excluded from training data and used as a test gallery. This ensures robustness to unseen faces for expression recognition. We have used the N-fold PI cross-validation scheme for CK+, MMI, ISED and DISFA datasets. Since all the expressions of every subject are not available in these datasets, the N-fold PI cross-validation is a more suitable scheme. However, in JAFFE and MUG and OULU-CASIA datasets, we have followed the leave-one-subject-out scheme as all the expressions for each subject is available.

The researchers in affective computing have adopted various dataset selection procedures and experimental setups. Therefore, it is difficult to make an appropriate comparison between the various published results. Also, most papers do not reveal the exact details of the PD or PI N-fold strategy. Therefore, to ensure a fair comparison, we have implemented the existing feature descriptors and incorporated them into our experimental setup. Moreover, in Table 2, we have tabulated the number of collected images in each class for all the datasets used in our experiments. The qualitative performances of the proposed method and the existing state-of-the-art techniques are clearly shown in Fig. 4. In the following subsections, we have presented the results of the proposed methods and compared with the state-of-the-art approaches on nine facial expression datasets.

5.1 Cohn–Kanade+ (CK+)

The CK+ [41, 42] database includes 593 image sequences of 123 different subjects. The subjects are of American, African–American, Asian and Latin origin. Each sequence starts with the neutral state and ends at the apex of an expression. The dataset provides frontal facial images of six expressions: anger, disgust, fear, happy, sad, and surprise. We selected three apex frames from each sequence to prepare the image set for an expression class. We also collected the neutral state images from the onset of the image sequences to create a neutral image set. Finally, we augmented a total of 1043 images: 132-anger, 180-disgust, 75-fear, 204-happy, 87-sad, 249-surprise, and 116-neutral. For PI setup, 1465 expression images (anger: 152, disgust: 190, fear: 144, happy: 247, sad: 181, surprise: 246, neutral: 305) were selected from the CK+ dataset.

In our experiments, the 10-fold PI cross-validation scheme was used in PI setup. In order to compare RADAP with deep learning techniques, we have trained the VGG16, VGG19 and ResNet50 networks over the CK+ dataset. The pre-trained weights from ImageNet were used as initial weight parameters while fine-tuning these networks. We have measured the performance of the FER system in terms of average recognition accuracy. In Table 3, we have shown a comprehensive comparative analysis of the proposed methods with state-of-the-art approaches including recent deep learning methods. From Table 3, it is evident that the proposed methods achieve superior recognition accuracy as compared to most of the existing feature descriptors. It also outperforms ResNet50 for both 6-class and 7-class recognition problem in PD setup. Moreover, it is clear that the proposed descriptors outperform LBP and LDN by 3.1 and 1.8% for a six-class problem and 6.4 and 3.7% for a seven-class problem, respectively, in PD experiments. Similarly, the proposed RADAP outperforms LBP and LDN by 0.66, 1.79 and 0.64, 1.25% for six-class and seven-class problem, respectively, in PI experiments.

Due to the random division of dataset into training and testing samples, the confusion matrix for the proposed method is computed by taking an average of all the confusion matrices generated at each iteration. We have shown the confusion matrices for six-class and seven-class problem in Figs. 5 and 6, respectively.

5.2 Japanese female facial expression (JAFFE)

The (JAFFE) [43] dataset contains facial images of ten Japanese females. The subjects posed for neutral and six basic facial expressions. Each expression set consists of an almost same

Table 2 Number of images used for different datasets in the experimental results

Emotion	CK+	JAFFE	MUG	MMI	COMBINED	OULU-VIS/NIR			ISED	DISFA	AFEW
exp. set	PD/PI	—	PD/PI	—	—	—	—	—	—	—	—
anger	132/152	30	260/220	69	635	240	240	240	—	35	1568
disgust	180/190	29	250/220	71	530	240	240	240	234	81	1148
fear	75/144	32	240/220	69	555	240	240	240	—	140	1193
happy	204/247	31	255/220	84	718	240	240	240	294	149	2137
sad	87/181	30	245/220	67	549	240	240	240	174	105	1782
surprise	249/246	30	255/220	75	609	240	240	240	189	79	930
neutral	116/305	30	255/220	126	605	240	240	240	—	141	—
total	1043/1465	212	1760/1540	561	4201	1680	1680	1680	891	730	8758

Exp. set: experimental setup. For the rest of the datasets (excluding CK+ & MUG), both PD and PI setup consists of the same number of images per class.

Table 3 Recognition accuracy comparison on CK+ dataset

Method	6EX(PD)	7EX(PD)	6EX(PI)	7EX(PI)
LBP [21]	93.5	89.0	89.97	83.96
two-phase [22]	88.2	79.5	79.54	72.85
LDP [28]	96.2	92.9	90.84	84.80
LDN [26]	94.8	91.7	88.84	83.35
LDTP [29]	95.3	91.9	89.60	83.08
LDTerP [27]	95.7	91.5	85.89	81.40
VGG16 [12]	96.7	95.2	91.31	88.18
VGG19 [12]	97.2	81.2	89.98	78.71
ResNet50 [13]	94.0	91.8	89.32	87.31
RADAP	96.2	94.7	88.48	83.72
XRADAP	96.6	93.8	84.64	79.13
ARADAP	96.2	93.8	87.84	83.77
DRADAP	96.0	95.4	90.63	84.60

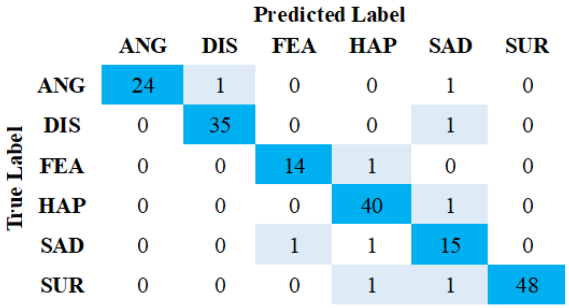

Fig. 5 *Confusion matrix of RADAP for six-class expression classification in CK+ dataset*

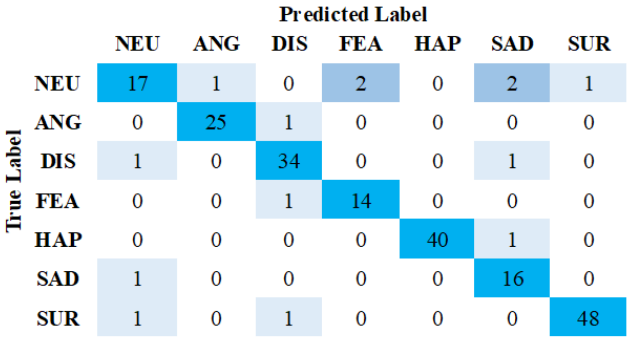

Fig. 6 *Confusion matrix of RADAP for seven-class expression classification in CK+ dataset*

number of images. The facial images were captured from the frontal view and hair of the female subjects were tied back to expose all the expressive zones of facial region. We have given the average recognition accuracy for six-class and seven-class problem over JAFFE dataset in Table 4. The proposed methods achieve better recognition accuracy as compared to state-of-the-art handcrafted approaches as well as some of the deep learning techniques. More specifically, the proposed methods attain recognition rate improvement of 9.8 and 8.3% for the six-class problem and 6.2 and 4.3% for the seven-class problem over LBP and LDN, respectively, in PD experiments. It also achieves 0.56, 0.56 and 2.55, 1.33% improvement over LBP and LDN for six-class and seven-class problems, respectively, in PI experiments.

5.3 Multimedia Understanding Group (MUG)

The MUG [44] dataset was created to overcome some of the shortcomings of other facial expression datasets such as low resolution, illumination variations, multiple takes per subject etc. It consists of image sequences from 86 subjects exhibiting various facial expressions. The subjects consisted of 35 females and 51 males of Caucasian origin aged between 20 and 35 years. The frontal view of the face was captured without any occlusion. In our

Table 4 Recognition accuracy comparison on JAFFE dataset

Method	6EX(PD)	7EX(PD)	6EX(PI)	7EX(PI)
LBP [21]	85.2	84.3	56.66	53.65
two-phase [22]	83.3	80.5	36.11	23.19
LDP [28]	85.0	80.5	52.77	52.10
LDN [26]	86.7	86.2	56.66	54.87
LDTP [29]	83.3	82.9	55.55	51.32
VGG16 [12]	61.1	73.8	68.29	59.47
VGG19 [12]	66.7	73.8	60.55	61.42
ResNet50 [13]	75.0	64.3	59.44	57.13
RADAP	95.0	88.1	55.56	55.71
XRADAP	88.4	88.1	51.67	51.43
ARADAP	91.1	88.1	56.11	56.20
DRADAP	93.9	90.5	57.22	55.71

Table 5 Recognition accuracy comparison on MUG dataset

Method	6EX(PD)	7EX(PD)	6EX(PI)	7EX(PI)
LBP [21]	97.5	97.0	82.65	76.16
two-phase [22]	97.7	97.2	74.46	70.12
LDP [28]	98.1	97.4	82.87	78.70
LDN [26]	98.8	97.9	81.96	77.85
LDTP [29]	98.1	98.0	82.04	78.70
LDTerP [27]	98.8	98.6	80.15	78.11
VGG16 [12]	99.6	98.8	85.14	84.67
VGG19 [12]	98.0	97.7	85.22	85.12
ResNet50 [13]	96.3	95.7	86.88	85.58
RADAP	99.3	98.5	82.65	78.57
XRADAP	98.9	98.2	80.68	75.26
ARADAP	99.2	98.2	83.41	79.68
DRADAP	99.0	98.7	83.48	80.26

experiment, the expression categories were formed by selecting five peak frames per sequence from 51 available subjects. We have a total of 1761 images (260-anger, 250-disgust, 241-fear, 255-happy, 245-sad, 255-surprise, and 255-neutral).

The recognition rates for the proposed methods and the other methods over MUG dataset is shown in Table 5 for six- and seven-class problem. From Table 5, we can conclude that the proposed methods outperform the state-of-the art techniques LBP, two-phase LBP, LDP, LDN, LDTP, and LDTerP in both six-class and seven-class expression recognition task. More specifically, we achieve the quantitative performance improvement over LBP and LDN by 1.8 and 0.5% for the six-class problem and 1.7 and 0.8% for the seven-class problem, respectively, in PD experiments. Likewise, from Fig. 4, it is evident that the qualitative performance of the proposed methods is better than other approaches. Moreover, it outperforms LBP and LDN by 0.83, 1.52 and 4.1, 2.41% for six- and seven-class problems, respectively, in PI experiments. All the methods achieve better recognition rates when compared with the results over CK+ and JAFFE datasets. The reason behind this is the sophisticated augmentation of the MUG dataset, which eliminates illumination variations and provides high-resolution images. Besides, the elicited expressions are relatively more distinguishable than other datasets.

5.4 MMI

The MMI expression database [45, 46] contains more than 2900 samples as videos and still images. Each session recording consists of the temporal sequence of facial expressions, beginning from a neutral state to a series of apex states and then back to the neutral state. The facial images are captured from the frontal and profiles views with multiple emotions. For persons wearing glasses; multiple sessions were recorded with and without their glasses on. In our experiments, we used the frontal face images from the 'videos with emotional label'. It consists of 236 sessions recorded

Table 6 Recognition accuracy comparison on MMI dataset

Method	6EX(PD)	7EX(PD)	6EX(PI)	7EX(PI)
LBP [21]	76.5	81.7	46.37	55.32
two-phase [22]	75.4	82.0	44.12	53.12
LDP [28]	80.5	84.0	45.70	55.82
LDN [26]	80.5	83.0	44.60	55.81
LDTP [29]	83.4	86.0	46.49	56.10
LDTerP [27]	80.6	80.0	43.60	54.47
VGG16 [12]	83.9	89.2	66.47	70.98
VGG19 [12]	81.6	83.9	63.68	69.95
ResNet50 [13]	71.2	83.9	59.37	63.49
RADAP	83.1	84.5	49.58	58.15
XRADAP	81.0	83.1	45.46	54.20
ARADAP	80.9	84.3	48.86	57.43
DRADAP	85.5	86.6	49.77	58.98

Table 7 Recognition accuracy comparison on OULU-CASIA (NIR) six-class dataset in PD setup

Method	6EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	97.6	97.2	97.2	97.3
two-phase [22]	94.3	94.1	95.2	94.5
LDP [28]	96.6	97.5	97.9	97.3
LDN [26]	98.3	98.1	98.5	98.3
LDTP [29]	98.1	98.0	98.2	98.1
LDTerP [27]	98.0	97.8	98.1	98.0
RADAP	99.0	98.2	98.9	98.7
XRADAP	97.4	97.3	97.2	97.3
ARADAP	98.6	98.3	99.0	98.6
DRADAP	98.7	98.5	98.5	98.6

for 32 subjects. We have selected the frame sequences of 30 subjects. Each sequence is labelled with one of the six (anger, sad, disgust, surprise, disgust, and happiness) to form a six-class dataset. Similarly, we have considered the neutral state by selecting onset frames from an image sequence and formed the seven-class dataset. We manually cropped the facial images only for MMI dataset. We have a total of 561 images (69-anger, 71-disgust, 69-fear, 84-happy, 67-sad, 75-surprise, and 126-neutral).

In this experiment, we carried out the person-dependent cross-validation process to evaluate the recognition accuracy of the FER system in PD experiments. Furthermore, we have used the ten-fold PI cross-validation scheme for PI experiments. The performance of the proposed descriptors and other existing techniques are measured on both the six-class and seven-class datasets. The recognition accuracy of these approaches is tabulated in Table 6. Notice that the results for the existing approaches are quite different from that presented in the original paper. This is due to the difference in collected data and experimental setup. From Table 6, it is clear that the proposed descriptors outperform the state-of-the-art feature descriptors and some of the deep learning approaches. More specifically, the proposed descriptors outperform LBP and LDN by 9 and 5% for the six-class problem and 4.9 and 3.6% for the seven-class problem, respectively. Furthermore, it outperforms LBP and LDN by 3.4, 5.17 and 3.66, 3.17% for six- and seven-class problems, respectively, in PI experiments. The qualitative performance comparison of the proposed methods and other approaches for a sample MMI expression image is shown in Fig. 4.

5.5 OULU-CASIA

The OULU-CASIA [47] database consists of six expressions (surprise, disgust, sad, fear, disgust, and happiness) from 80 subjects. These subjects are in the age group of 23–58 years. The subjects were instructed to face the frontal view of the camera attached to a computer monitor. The expressions were captured

Table 8 Recognition accuracy comparison on OULU-CASIA (NIR) seven-class dataset in PD setup

Method	7EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	96.4	96.9	95.9	96.4
two-phase [22]	93.0	92.3	91.3	92.2
LDP [28]	96.0	97.7	97.7	97.1
LDN [26]	96.7	98.1	98.0	97.6
LDTP [29]	97.8	97.7	97.1	97.5
LDTerP [27]	97.7	96.6	98.2	97.5
RADAP	98.7	98.1	97.2	98
XRADAP	97.0	97.1	97.2	97.1
ARADAP	98.1	99.2	98.3	98.5
DRADAP	97.7	98.2	98.3	98.1

Table 9 Recognition accuracy comparison on OULU-CASIA (NIR) six-class dataset in PI setup

Method	6EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	64.09	67.09	63.26	64.81
two-phase [22]	47.36	48.51	44.93	46.93
LDP [28]	64.09	59.44	63.05	62.19
LDN [26]	63.05	65.48	64.65	64.39
LDTP [29]	61.11	67.77	62.36	63.74
LDTerP [27]	56.04	57.08	51.31	54.81
RADAP	64.65	65.69	65.76	65.37
XRADAP	60.10	63.40	62.71	62.06
ARADAP	63.88	66.11	64.72	64.90
DRADAP	64.65	67.29	65.56	65.83

using near infrared (NIR) and visible (VIS) cameras simultaneously. Furthermore, all expressions are recorded in three different illumination conditions: strong, dark, and weak. In the strong category, appropriate lighting is used. In the weak category, only the computer display is on, whereas, dark category means near darkness. For each illumination category, 480 video sequences were recorded. So, totally there are 2880 video sequences present in the dataset. In our experiments, we have selected three peak frames from each expression and arranged the dataset for each illumination category. The images for a neutral state were collected from the onset of each recording session. In our work, the following set of expressions are collected for each illumination condition: 240-anger, 240-fear, 240-happy, 240-sad, 240-disgust, 240-surprise, and 240-neutral. The same is applied for both NIR and VIS camera image sequences.

We have calculated the recognition accuracy of the proposed descriptors and other existing techniques in both OULU-CASIA NIR and OULU-CASIA VIS datasets. In Tables 7–10, we have shown the results for the six-class and seven-class recognition problem in the OULU-CASIA NIR dataset. It is clear from Tables 7–10 that the proposed descriptors outperform other feature descriptors. More specifically, the proposed method achieves 1.4 and 0.4% avg. recognition accuracy improvement over LBP and LDN for the six-class problem in OULU-CASIA NIR dataset for the PD setup. Similarly, it achieves 2.1 and 0.9 improvements over LBP and LDP for the seven-class problem in the same dataset. Moreover, the comparative results for PI experiments over NIR dataset as shown in Tables 9 and 10, further demonstrate the effectiveness of the proposed RADAP descriptors. The proposed method attains 1.02 and 1.44% performance improvement over LBP and LDN in the six-class recognition problem. It also outperforms LBP and LDN by 1.69 and 2.53% in seven-class expression recognition.

Furthermore, we have computed the results over OULU-CASIA VIS dataset and showed the results for six-class and seven-class emotions in Tables 11–14. In PD setup, the proposed method outperforms LBP and LDN by 2.1 and 1.4% for the six-class

Table 10 Recognition accuracy comparison on OULU-CASIA (NIR) seven-class dataset in PI setup

Method	7EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	63.63	65.71	61.84	63.72
two-phase [22]	45.41	46.30	44.58	45.43
LDP [28]	62.97	65.22	62.32	63.50
LDN [26]	63.21	64.82	60.61	62.88
LDTP [29]	61.13	66.72	62.14	63.33
LDTerP [27]	40.65	55.83	51.96	49.48
RADAP	65.35	66.19	64.70	65.41
XRADAP	61.07	63.57	62.91	62.52
ARADAP	63.86	66.96	65.05	65.29
DRADAP	64.22	66.78	65.00	65.33

Table 11 Recognition accuracy comparison on OULU-CASIA (VIS) six-class dataset in PD setup

Method	6EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	94.1	96.3	96.1	95.5
two-phase [22]	80.3	87.8	90.0	86.0
LDP [28]	92.7	98.4	97.2	96.1
LDN [26]	94.3	98.5	96.0	96.2
LDTP [29]	90.3	98.5	96.6	95.1
LDTerP [27]	93.9	98.3	97.2	96.4
RADAP	95.9	99.0	98.0	97.6
XRADAP	93.3	97.7	96.0	96.2
ARADAP	94.9	98.0	97.3	96.7
DRADAP	93.9	98.3	97.2	96.5

Table 12 Recognition accuracy comparison on OULU-CASIA (VIS) seven-class dataset in PD setup

Method	7EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	90.1	93.3	94.1	92.5
two-phase [22]	86.2	87.0	89.4	87.5
LDP [28]	94.3	98.0	96.3	96.2
LDN [26]	95.3	97.8	96.7	96.6
LDTP [29]	95.0	98.3	96.7	96.7
LDTerP [27]	92.4	98.8	96.8	96.0
RADAP	97.1	98.8	97.3	97.7
XRADAP	94.3	97.9	96.0	96.1
ARADAP	94.1	98.3	97.1	96.5
DRADAP	95.0	98.3	96.9	96.7

problem and 5.2 and 1.1% for the seven-class problem, respectively. Moreover, in the PI setup, the proposed method outperforms LBP and LDN by 0.05 and 1.46% for the six-class problem and 2.62 and 1.41% for the seven-class problem, respectively.

The qualitative responses of the proposed methods are evaluated by taking a sample facial image from the OULU-CASIA dataset as shown in Fig. 4. From Fig. 4 and Tables 7–14, it is clear that the proposed descriptors robustly recognise the facial features in the presence of illumination variation.

5.6 Combined dataset

One of the drawbacks in selecting the three most expressive frames from an image sequence is that there is a lot of similarity between those images. Thus, a random partition may sometimes result in the presence of similar expression images in training and testing sets. This may influence the task of classification. However, without such a setup, the training dataset will become very small and thus

Table 13 Recognition accuracy comparison on OULU-CASIA (VIS) six-class dataset in PI setup

Method	6EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	55.90	75.34	57.70	62.98
two-phase [22]	31.04	55.83	38.47	41.78
LDP [28]	46.90	72.43	56.45	58.60
LDN [26]	53.81	72.29	58.61	61.57
LDTP [29]	42.36	72.29	55.62	56.75
LDTerP [27]	43.26	68.54	52.70	54.83
RADAP	52.64	75.83	60.63	63.03
XRADAP	46.59	72.78	56.74	58.70
ARADAP	48.75	75.90	59.65	61.44
DRADAP	49.86	75.69	60.90	62.15

Table 14 Recognition accuracy comparison on OULU-CASIA (VIS) seven-class dataset in PI setup

Method	7EX average accuracy			
	Dark	Strong	Weak	Avg.
LBP [21]	57.38	65.71	57.32	60.13
two-phase [22]	28.09	53.98	39.22	40.43
LDP [28]	53.86	69.16	54.88	59.30
LDN [26]	53.86	70.89	59.28	61.34
LDTP [29]	42.08	68.86	54.94	55.29
LDTerP [27]	40.65	64.53	48.75	51.31
RADAP	52.97	74.11	61.07	62.72
XRADAP	47.74	70.41	56.19	58.11
ARADAP	50.83	72.32	60.29	61.15
DRADAP	52.97	74.34	60.95	62.75

Table 15 Recognition accuracy comparison on the combined (CK+, JAFFE, MUG, MMI & GEMEP-FERA) dataset

Method	6EX(PD)	7EX(PD)
LBP [21]	90.1	89.0
two-phase [22]	88.4	86.2
LDP [28]	92.5	90.2
LDN [26]	93.8	93.5
LDTP [29]	94.2	91.8
VGG16 [12]	92.5	94.5
VGG19 [12]	93.8	94.1
ResNet50 [13]	89.4	88.5
RADAP	95.0	94.0
XRADAP	93.9	92.2
ARADAP	95.2	93.7
DRADAP	94.8	93.0

lead to improperly trained SVM models. Therefore, to further validate the effectiveness of the proposed methods, we amalgamate the CK+ [41, 42], JAFFE [43], MUG [44], MMI [45, 46] and GEMEP-FERA [53, 54] datasets and augment a larger pool of data for training and testing. The combined dataset contains a total of 4021 images: 635-anger, 530-disgust, 555-fear, 718-happy, 549-sad, 609-surprise, and 605-neutral. The SVM model was trained and tested over 3217 and 804 images, respectively.

The performance measure of the proposed methods and other approaches over the combined dataset is shown in Table 15. The proposed methods surpass the results of the state-of-the-art handcrafted techniques in PD experiments. It obtains 5.1, 1.4 and 5, 0.5% improvement over LBP and LDN for six-class and seven-class problem, respectively. Moreover, it achieves better recognition accuracy as compared to some of the deep learning techniques. Thus, this experiment further increases the credibility of the proposed methods for effective FER.

Table 16 Recognition accuracy comparison on ISED dataset

Method	4EX(PD)	4EX(PI)
LBP [21]	88.9	63.29
two-phase [22]	78.4	58.57
LDP [28]	88.3	64.42
LDN [26]	90.6	64.19
LDTP [29]	91.0	64.30
LDTerP [27]	88.0	62.86
VGG16 [12]	95.7	73.09
VGG19 [12]	92.9	70.24
ResNet50 [13]	90.2	68.10
RADAP	91.5	67.05
XRADAP	91.2	63.02
ARADAP	93.5	66.02
DRADAP	92.2	68.81

Table 17 Recognition accuracy comparison on DISFA dataset

Method	6EX(PD)	7EX(PD)	6EX(PI)	7EX(PI)
LBP [21]	91.8	92.7	53.84	53.58
two-phase [22]	91.0	92.8	53.30	51.61
LDP [28]	91.5	94.1	52.00	49.65
LDN [26]	90.7	93.0	56.39	51.59
LDTP [29]	92.2	93.8	58.30	55.12
VGG16 [12]	89.2	83.9	56.76	57.42
VGG19 [12]	83.9	88.3	59.98	53.66
ResNet50 [13]	83.9	71.2	62.48	54.01
RADAP	93.2	95.1	62.38	59.71
XRADAP	92.4	94.7	61.17	57.21
ARADAP	93.8	95.1	61.73	59.94
DRADAP	94.9	95.3	62.80	60.55

5.7 Indian spontaneous expression database (ISED)

Moreover, we evaluated the spontaneous expression recognition problem in ISED [48]. In ISED, the facial expressions of induced spontaneous emotion of the participants were collected. These sessions were self-rated by the subjects based on the experienced emotion. Furthermore, these annotations were validated by trained decoders based on the nature of stimuli and self-reported emotions. The dataset contains near frontal facial recordings of 50 subjects. The head movements were allowed in all directions. It covers four emotions: happiness (227 clips), disgust (80 clips), sadness (48 clips), and surprise (73 clips). We re-arranged the dataset by selecting the peak frames as mentioned in previous experiments. Finally, we have four subsets: happy-294, sad-174, surprise-189, and disgust-234. We selected a single peak image from each session of happy emotions to avoid data imbalance for classification. For the PI setup, we have employed the ten-fold PI cross-validation scheme.

The experimental results of the proposed methods and existing state-of-the-art approaches are given in Table 16. The proposed methods perform better than state-of-the-art feature descriptors and some of the deep learning approaches. Particularly, it outperforms LBP and LDN by 4.6 and 2.9% in PD experiments. Moreover, in the PI setup, the proposed RADAP descriptors outperform LBP and LDN by 5.52 and 4.62%, respectively. This proves the effectiveness of our work in spontaneous expression recognition problems as well.

5.8 DISFA

DISFA [49] dataset contains around 89,000 video frames of 27 subjects (15 male and 12 female) aged between 18 and 50 years. Each video was captured while subjects were watching different video clips without their awareness to draw out spontaneous expressions. The video frames have been manually coded with different AU intensities according to FACs. In our setup, we selected four to five most expressive frames from each video sequence and collected 730 images. The recognition rate of the proposed method and existing methods are tabulated in Table 17. It is clear that the proposed methods outperform the state-of-the-art feature descriptors and some of the deep learning approaches. More specifically, the proposed method improves the recognition accuracy rate by 3.1, 4.2 and 2.6, 2.3% over LBP and LDTP, respectively, for six-class and seven-class problems. Furthermore, in PI experiments, the RADAP descriptors outperform LBP and LDN by 8.96, 6.41 and 6.97, 8.96% for the six-class and seven-class problems, respectively.

Table 18 Recognition accuracy comparison on AFEW dataset

Method	6EX(PD)
LBP [21]	80.5
two-phase [22]	57.0
LDP [28]	77.2
LDN [26]	81.5
LDTP [29]	77.8
LDTerP [27]	83.9
VGG16 [12]	97.0
VGG19 [12]	97.5
ResNet50 [13]	93.9
RADAP	86.7
XRADAP	82.6
ARADAP	88.7
DRADAP	87.4

5.9 AFEW

AFEW [50, 51] contains 1156 video clips of different emotions, which were extracted from the movie scenes. The dataset presents various real-world constraints such as spontaneous facial expressions, head pose variations, occlusions, variable face resolution, and illumination changes. We have selected 10–15 peak frames from each video clip and augmented a static facial expression dataset consisting of 8758 images. From Table 18, we can see that the proposed method outperforms the state-of-the-art handcrafted approaches in FER. More specifically, it yields 8.2 and 4.8% more accuracy as compared to LBP and LDTerP, respectively, for six-class expression recognition in the PD setup.

5.10 Computation time

We also compare the computation time performance of the proposed methods and existing state-of-the-art approaches.

The computation time includes loading the input, feature extraction and output label prediction using the trained SVM/deep learning model. The experiments were conducted over a system with following configurations: Operating System: Ubuntu 16.04, Processor: Xeon E5-2630 v4, 2.20 GHz x40, RAM: 64 GB, Tool: MATLAB. In Table 19, we present the comparative performance analysis of the proposed and existing descriptors in terms of computation time.

6 Conclusion

In this study, we proposed a novel feature descriptor RADAP by developing an adaptive global threshold generation technique to encode features at multiple distances within the local neighbourhood. It captures both the local and globally invariant features in the local region and therefore is robust to noise and uneven illumination changes. Furthermore, we extended our work to encode pairwise co-occurrence and fused cross-co-occurrence patterns in the RADAP features. The XRADAP improves robustness to intra-class variations within the same emotion. The ARADAP and DRADAP extract more stable and illumination

Table 19 Computation time comparison of the proposed methods and other state-of-the-art approaches

Method	Com. time, s	Method	Com. time, s
LBP [21]	0.05498	VGG19 [12]	1.44195
LDP [28]	0.10544	ResNet50 [13]	1.34389
LDN [26]	0.07962	RADAP	0.13666
LDTP [29]	0.08553	XRADAP	0.14905
LDTerP [27]	0.66089	ARADAP	0.14476
two-phase [22]	0.04651	DRADAP	0.76357
VGG16 [12]	1.41934	—	—

Com.: Computation, s: seconds.

invariant features for better results in a real world environment. The experiments were conducted in both PD and PI setup. From the experimental results, it is evident that the proposed RADAP descriptors outperform existing state-of-the-art techniques in nine different facial expression datasets. In the future, the concept of cross-distance co-occurrence can be applied to another feature descriptor as well and their impact in different computer vision applications can be studied.

7 Acknowledgments

This work was supported by the Science and Engineering Research Board (under the Department of Science and Technology, Govt. of India) project #SERB/F/9507/2017. The authors would also like to thank the members of Vision Intelligence Lab for their valuable support.

8 References

[1] Donato, G., Barlett, M.S., Hager, J.C., *et al.*: 'Classifying facial actions', *IEEE Trans. Pattern Anal. Mach. Intell.*, 1999, **21**, (10), pp. 974–989

[2] Ekman, P., Friesen, W.V.: 'Constants across cultures in the face and emotion', *J. Personal. Social Psychol.*, 1971, **17**, (2), p. 124

[3] Ekman, P.: 'An argument for basic emotions', *Cogn. Emot.*, 1992, **6**, (3–4), pp. 169–200

[4] Corneanu, C.A., Simón, M.O., Cohn, J.F., *et al.*: 'Survey on RGB, 3D, thermal, and multimodal approaches for facial expression recognition: history, trends, and affect-related applications', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016, **38**, (8), pp. 1548–1568

[5] Burkert, P., Trier, F., Afzal, M.Z., *et al.*: 'DeXpression: deep convolutional neural network for expression recognition', arXiv preprint arXiv:1509.05371, 2015

[6] Mollahosseini, A., Chan, D., Mahoor, M.H.: 'Going deeper in facial expression recognition using deep neural networks'. Proc. IEEE Winter Conf. Applications Computer Vision, 2016, pp. 1–10

[7] Barsoum, E., Zhang, C., Ferrer, C.C., *et al.*: 'Training deep networks for facial expression recognition with crowd-sourced label distribution'. Proc. 18th ACM Int. Conf. on Multimodal Interaction, 2016, pp. 279–283

[8] Hasani, B., Mahoor, M.H.: 'Facial expression recognition using enhanced deep 3D convolutional neural networks'. Proc. IEEE Conf. Computer Vision Pattern Recognition Workshops, 2017, pp. 2278–2288

[9] Ding, H., Zhou, S.K., Chellappa, R.: 'FaceNet2ExpNet: regularizing a deep face recognition net for expression recognition'. Proc. 12th IEEE Int. Conf. on Automatic Face Gesture Recognition, 2017, pp. 118–126

[10] Pons, G., Masip, D.: 'Supervised committee of convolutional neural networks in automated facial expression analysis', *IEEE Trans. Affect. Comput.*, 2017, **9**, (3), pp. 343–350

[11] Kim, B.K., Roh, J., Dong, S.Y., *et al.*: 'Hierarchical committee of deep convolutional neural networks for robust facial expression recognition', *J. Multimod. User Interfaces*, 2016, **10**, (2), pp. 1–17

[12] Simonyan, K., Zisserman, A.: 'Very deep convolutional networks for large-scale image recognition', arXiv preprint arXiv:1409.1556, 2014

[13] He, K., Zhang, X., Ren, S., *et al.*: 'Deep residual learning for image recognition'. Proc. IEEE Conf. on Computer Vision Pattern Recognition, 2016, pp. 770–778

[14] Jung, H., Lee, S., Park, S., *et al.*: 'Deep temporal appearance-geometry network for facial expression recognition', arXiv preprint arXiv:1503.01532, 2015

[15] Jung, H., Lee, S., Yim, J., *et al.*: 'Joint fine-tuning in deep neural networks for facial expression recognition'. Proc. IEEE Int. Conf. on Computer Vision, 2015, pp. 2983–2991

[16] Zhang, K., Huang, Y., Du, Y., *et al.*: 'Facial expression recognition based on deep evolutional spatial-temporal networks', *IEEE Trans. Image Proc.*, 2017, **26**, (9), pp. 4193–4203

[17] Kim, Y., Yoo, B., Kwak, Y., *et al.*: 'Deep generative-contrastive networks for facial expression recognition', arXiv preprint arXiv:1703.07140, 2017

[18] Sariyanidi, E., Gunes, H., Cavallaro, A.: 'Automatic analysis of facial affect: A survey of registration, representation, and recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (6), pp. 1113–1133

[19] Pantic, M., Patras, I.: 'Dynamics of facial expression: recognition of facial actions and their temporal segments from face profile image sequences', *IEEE Trans. Syst. Man Cybern. B, Cybern.*, 2006, **36**, (2), pp. 433–449

[20] Sebe, N., Lew, M.S., Sun, Y., *et al.*: 'Authentic facial expression analysis', *Image Vis. Comput.*, 2007, **25**, (12), pp. 1856–1863

[21] Shan, C., Gong, S., McOwan, P.W.: 'Facial expression recognition based on local binary patterns: a comprehensive study', *Image Vis. Comput.*, 2009, **27**, (6), pp. 803–816

[22] Lai, C.C., Ko, C.H.: 'Facial expression recognition based on two-stage features extraction', *Optik*, 2014, **125**, (22), pp. 6678–6680

[23] Dhall, A., Asthana, A., Goecke, R., *et al.*: 'Emotion recognition using PHOG and LPQ features'. Proc. IEEE Int. Conf. on Automatic Face Gesture Recognition and Workshops, 2011, pp. 878–883

[24] Murala, S., Wu, Q.M.J.: 'Local mesh patterns versus local binary patterns: biomedical image indexing and retrieval', *IEEE J. Biomed. Health Informatics*, 2014, **18**, (3), pp. 929–938

[25] Yang, S., Bhanu, B.: 'Facial expression recognition using emotion avatar image'. Proc. IEEE Int. Conf. on Automatic Face Gesture Recognition and Workshops, 2011, pp. 866–871

[26] Rivera, A.R., Castillo, J.R., Chae, O.O.: 'Local directional number pattern for face analysis: face and expression recognition', *IEEE Trans. Image Process.*, 2013, **22**, (5), pp. 1740–1752

[27] Ryu, B., Rivera, A.R., Kim, J., *et al.*: 'Local directional ternary pattern for facial expression recognition', *IEEE Trans. Image Process.*, 2017, **26**, (12), pp. 6006–6018

[28] Jabid, T., Kabir, M., Chae, O.O.: 'Robust facial expression recognition based on local directional pattern', *ETRI J.*, 2010, **32**, (5), pp. 784–794

[29] Rivera, A.R., Castillo, J.R., Chae, O.O.: 'Local directional texture pattern image descriptor', *Pattern Recognit. Lett.*, 2015, **51**, pp. 94–100

[30] Zhang, H., Fritts, J.E., Goldman, S.A.: 'Image segmentation evaluation: A survey of unsupervised methods', *Comput. Vis. Image Understand.*, 2008, **110**, (2), pp. 260–280

[31] Zhang, Z., Lyons, M., Schuster, M., *et al.*: 'Comparison between geometric-based and Gabor-wavelets-based facial expression recognition using multi-layer perceptron'. Proc. IEEE Int. Conf. on Automatic Face Gesture Recognition, 1998, pp. 454–459

[32] Valstar, M.F., Patras, I., Pantic, M.: 'Facial action unit detection using probabilistic actively learned support vector machines on tracked facial point data'. Proc. IEEE Conf. on Computer Vision Pattern Recognition Workshops, 2005, pp. 76–76

[33] Lee, S.H., Plataniotis, K.N.K., Ro, Y.M.: 'Intra-class variation reduction using training expression images for sparse representation based facial expression recognition', *IEEE Trans. Affect. Comput.*, 2014, **5**, (3), pp. 340–351

[34] Mohammadzade, H., Hatzinakos, D.: 'Projection into expression subspaces for face recognition from single sample per person', *IEEE Trans. Affect. Comput.*, 2013, **4**, (1), pp. 69–82

[35] Zhang, L., Tjondronegoro, D.: 'Facial expression recognition using facial movement features', *IEEE Trans. Affect. Comput.*, 2011, **2**, (4), pp. 219–229

[36] Happy, S.L., Routray, A.: 'Automatic facial expression recognition using features of salient facial patches', *IEEE Trans. Affect. Comput.*, 2015, **6**, (1), pp. 1–12

[37] Zhang, T., Zheng, W., Cui, Z., *et al.*: 'A deep neural network-driven feature learning method for multi-view facial expression recognition', *IEEE Trans. Multimedia*, 2016, **18**, (12), pp. 2528–2536

[38] Zheng, W., Zong, Y., Zhou, X., *et al.*: 'Cross-domain color facial expression recognition using transductive transfer subspace learning', *IEEE Trans. Affect. Comput.*, 2018, **9**, (1), pp. 21–37

[39] Rudovic, O., Pantic, M., Patras, I.: 'Coupled Gaussian processes for pose-invariant facial expression recognition', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (6), pp. 1357–1369

[40] Burges, C.J.C.: 'A tutorial on support vector machines for pattern recognition', *Data Min. Knowl. Discov.*, 1998, **2**, (2), pp. 121–167

[41] Kanade, T., Cohn, J.F., Tian, Y.: 'Comprehensive database for facial expression analysis'. Proc. 4th IEEE Int. Conf. on Automatic Face Gesture Recognition, 2000, pp. 46–53

[42] Lucey, P., Cohn, J.F., Kanade, T., *et al.*: 'The extended Cohn–Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression'. Proc. IEEE Conf. on Computer Vision Pattern Recognition Workshops, 2010, pp. 94–101

[43] Lyons, M., Akamatsu, S., Kamachi, M., *et al.*: 'Coding facial expressions with Gabor wavelets'. Proc. Third IEEE Int. Conf. on Automatic Face Gesture Recognition, 1998, pp. 200–205

[44] Aifanti, N., Papachristou, C., Delopoulos, A.: 'The MUG facial expression database'. Proc. 11th Int. Workshop on Image Analysis for Multimedia Interactive Services, 2010, pp. 1–4

[45] Pantic, M., Valstar, M., Rademaker, R., *et al.*: 'Web-based database for facial expression analysis'. Proc. IEEE Int. Conf. on Multimedia Expo, 2005

[46] Valstar, M., Pantic, M.: 'Induced disgust, happiness and surprise: an addition to the mmi facial expression database'. Proc. 3rd Int. Workshop Emotion (Satellite of LREC), 2010, pp. 65–70

[47] Zhao, G., Huang, X., Taini, M., *et al.*: 'Facial expression recognition from near-infrared videos', *Image Vis. Comput.*, 2011, **29**, (9), pp. 607–619

[48] Happy, S.L., Patnaik, P., Routray, A., *et al.*: 'The Indian spontaneous expression database for emotion recognition', *IEEE Trans. Affect. Comput.*, 2017, **8**, (1), pp. 131–142

[49] Mavadati, S.M., Mahoor, M.H., Bartlett, K., *et al.*: 'Disfa: a spontaneous facial action intensity database', *IEEE Trans. Affect. Comput.*, 2013, **4**, (2), pp. 151–160

[50] Dhall, A., Goecke, R., Lucey, S., *et al.*: 'Collecting large, richly annotated facial-expression databases from movies', *IEEE Multimed.*, 2012, **19**, (3), pp. 34–41

[51] Dhall, A., Goecke, R., Ghosh, S., *et al.*: 'From individual to group-level emotion recognition: EmotiW 5.0.'. Proc. 19th ACM Int. Conf. on Multimodal Interaction, 2017, pp. 524–528
[52] Viola, P., Jones, M.J.: 'Robust real-time face detection', *Int. J. Comput. Vis.*, 2004, **57**, (2), pp. 137–154
[53] Valstar, M.F., Jiang, B., Mehu, M., *et al.*: 'The first facial expression recognition and analysis challenge'. IEEE Int. Conf. on Automatic Face Gesture Recognition Workshops, 2011, pp. 921–926
[54] Bänziger, T., Mortillaro, M., Scherer, K.R.: 'Introducing the Geneva Multimodal Expression corpus for experimental research on emotion perception', *Emotion*, 2012, **12**, (5), p. 1161