Spontaneous Expression Recognition using Universal Attribute Model

Article *in* IEEE Transactions on Image Processing · July 2018

DOI: 10.1109/TIP.2018.2856373

3 authors:

Nazil Perveen
Indian Institute of Technology Hyderabad
6 PUBLICATIONS **27** CITATIONS
SEE PROFILE

Debaditya Roy
Nihon University
14 PUBLICATIONS **47** CITATIONS
SEE PROFILE

Krishna Mohan Chalavadi
Indian Institute of Technology Hyderabad
57 PUBLICATIONS **351** CITATIONS
SEE PROFILE

Spontaneous Expression Recognition using Universal Attribute Model

Nazil Perveen, Debaditya Roy, and C. Krishna Mohan

***Abstract*—Spontaneous expression recognition refers to recognizing non-posed human expressions. In literature, most of the existing approaches for expression recognition mainly rely on manual annotations by experts, which is both time-consuming and difficult to obtain. Hence, we propose an unsupervised framework for spontaneous expression recognition that preserves discriminative information for the videos of each expression without using annotations. Initially, a large Gaussian mixture model called universal attribute model (UAM) is trained to learn the attributes of various expressions implicitly. Attributes are the movements of various facial muscles that are combined to form a particular facial expression. Then a concatenated mean vector called the super expression-vector (SEV) is formed by using a maximum *a posteriori* adaptation of the UAM means for each expression clip. This SEV contains attributes from all the expressions resulting in a high dimensional representation. To retain only the attributes of that particular expression clip, the SEV is decomposed using factor analysis to produce a low-dimensional expression-vector. This procedure does not require any class labels and produces expression-vectors that are distinct for each expression irrespective of high inter-actor variability present in spontaneous expressions. On spontaneous expression datasets like BP4D and AFEW, we demonstrate that expression-vector achieves better performance than state-of-the-art techniques. Further, we also show that UAM trained on a constrained dataset can be effectively used to recognize expressions in unconstrained expression videos.**

***Index Terms*—Expression recognition, feature extraction, universal attribute model, map adaptation, factor analysis, Gaussian mixture model.**

I. Introduction

Automated facial expression recognition (AFER) has immense potential for application in neuromarketing, psychological treatment, interrogation simulators, robotics, real-time gaming, recommendation systems and so on. One of the biggest challenges in AFER is to recognize emotions such as anger, disgust, fear, happiness, sadness, and surprise in natural human expressions known as spontaneous expression recognnition (SER) [1]. The area of spontaneous expression recognition (SER) has witnessed ample research in the past decade to differentiate it from deliberate facial expressions that are easier to recognize [2]–[7] . Attempts have been made to recognize genuine smiles [8], posed versus fake expressions [9], pain, frustration, and fatigue [10] in spontaneous environments. These attempts involve using 3D information for view-independent analysis [11], and using thermal and audiovisual information, mainly to suit the needs of better facial expression recognition in the wild [12]. Also, spontaneous expressions have been analyzed intensively using RGB characteristics, relative speed, and timing information [13]–[15]. For evaluation, multiple databases like CK+ [16], BU-3DFE [17], BU-4DFE [18], and BP4D [19] consist of spontaneous expressions captured in a controlled environment. More challenging databases like acted facial expression in the wild (AFEW) [20] and affective MIT facial expression database [21] capture spontaneous expressions in an unconstrained environment (also known as in-the-wild).

One of the most popular methods for recognizing spontaneous facial behaviour has been the use of facial action units (AU) where each AU is defined as a contraction or relaxation of one or more facial muscles [22], [23]. As automatic AU recognition is challenging, some approaches have aimed to capture facial movements directly using spatio-temporal representations like local binary patterns - three orthogonal planes (LBP-TOP) [24], histogram of gradients (HOG)/ histogram of optical flow (HOF) [25], HOG3D [26], and 3D scale-invariant feature transform (SIFT) [27]. As these features are captured locally, they are used to perform expression analysis on the part of the face that is termed as micro-expression recognition [28].

One such approach from Yong et al [29] recognizes spontaneous micro-expressions using main directional mean optical flow (MDMO) feature representation that is computed from the local optical flow and spatial location information. In MDMO, only the highest optical flow in a particular region of interest is considered whereas in HOF, flow in all the directions are weighted. An SVM trained with MDMO features was shown to work better than HOF and LBP-TOP for recognizing micro-expressions. However, local representations often suffer from noise and spatio-temporal sensitivity. To obtain a more robust mid-level representation, expressionlets were proposed in [30]. At first, a Gaussian mixture model called universal manifold model (UMM) was learned, where each mixture captures the spatial and temporal correspondence in each local spatio-temporal manifold using HOG3D and 3DSIFT features. For each spatio-temporal manifold (STM), top T local features are chosen per mixture of the universal manifold model (UMM) using posterior probability. Every video is considered as an unaligned STM. The top T features per mixture are concatenated to form an aligned STM of dimension $C \times T \times d$, where C is the number of mixtures in UMM, and d is the dimension of the local feature. The means of top T features per mixture are calculated and corresponding UMM means are subtracted to get centered covariances. These concatenated covariances form expressionlets of dimension $C \times d$. Each

Nazil Perveen, Debaditya Roy and C. Krishna Mohan are with the *Visual Learning and Intelligence Group (VIGIL)*, Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad, India, (e-mail: {cs14resch11006,cs13p1001,ckm}@iith.ac.in).

expression video contains feature vector that represents the features that best matches each of the mixtures.

While most of the above methods work in constrained environments, the captured views of the subject in the wild are often either occluded or profiled. In such cases, spatial and appearance similarity is hard to establish, which can lead to problems in learning a view-invariant representation. Further, a classifier or discriminant has to be trained, which means that clear discrimination cannot be achieved inherently with local or mid-level representations. Also, the learned representations are high-dimensional that induces the curse of dimensionality in the case of a vast number of videos. In the proposed work, we provide an actor-invariant, low-dimensional, and high-level representation for spontaneous expression recognition in the wild which we call the *expression-vector*. We explore histogram of oriented gradients (HOF) and motion boundary histogram information (MBH) [31] for capturing facial muscle movements that we term expression *attributes*.

The aim of the proposed method is to capture the attributes of all expressions in a single model. Hence, a Gaussian mixture model known as a universal attribute model (UAM) is learned separately for the HOF and MBH descriptors. Then using the mean adaptation of the UAM mixtures, a super vector representation called super expression vector (SEV) is obtained for every expression clip. Since the SEV contains the information from all the attributes, it is intrinsically high-dimensional ($num\_mixtures \times feature\_dim$). To capture only the essential attributes, the SEV is decomposed using factor analysis which yields a low-dimensional (200) expression-vector representation. We show the efficacy of the proposed approach on BP4D and AFEW datasets. To summarize, the contributions of the proposed method are listed as follows:

- A universal attribute model (UAM) for learning expression attributes implicitly and subsequent extraction of a discriminative low-dimensional representation from the UAM called expression-vector.
- Spontaneous expression recognition of unconstrained video clips using a UAM trained on a constrained dataset.

The rest of the paper is organized as follows. In Section II, we discuss existing approaches for spontaneous expression recognition. Then in section III, unsupervised expression-vector extraction is described in detail with UAM construction, SEV generation, and low-dimensional decomposition of SEV. Section IV introduces the various expression-vector scoring mechanisms used in this work. The experimental results on benchmark databases are discussed in Section V and the conclusions are presented in Section VI.

II. Related Work

To accurately recognize expressions, representation of facial expression dynamics is of utmost importance. Based on the extraction of features, most of the methods in literature perform either explicit modelling or implicit modelling of expression dynamics. Explicit modelling methods correspond to hand-coded extraction of facial features like HOG $\text{hog}_1, SIFT1001$[32], $Gaborwavelets$ 1001[33], $Gaussiantransformations$ 1001[34], $localphasequantization(LPQ)$ 1001[35], $andsoon.Morerecently, domain-specificdescriptorslikethepyramidofHOG(PHOG)$ 1001[36], $localbinarypattern$ 1001[37], $etc.havegainedimportance.Whilesuchdescriptorsaremeant$ $TOP1001$[38]$andLBP-TOP$ 1001[39]$alsorecordtemporaldynamicsthatcanprovidediscriminativei$ 1001[40], $bothhand-codedfeaturesforspatialrepresentationandLBP$ 1001[41]$whereinsteadofextractingthefeaturesoverallttheattributes,$ $basedselectionscheme, specificfacialpatchescorrespondingtoeachex$

Videos captured in unconstrained environments mostly contain spontaneous facial expressions like Binghamton Pittsburgh 4D (BP4D) and acted facial expression in the wild (AFEW). In such environments, cross-database and subject independent evaluations have been used because: 1) models can be trained on constrained datasets and tested on unconstrained datasets [42] and 2) collecting sufficient samples for each subject in unconstrained environments is both challenging and time-consuming. Such evaluations give better insights into handling challenges like illumination, pose, and alignment across multiple datasets that help in testing the generalization capability of the learned model. Moreover, subject-independent evaluation has also been applied in online facial expression recognition [43] and in some scenarios like lie detection, where the subjects are not readily available [44].

III. Proposed Approach

Following the literature presented in the previous section, we use explicit features for building the universal attribute model (UAM). Figure 1 presents the block diagram for the proposed method. The various stages of the proposed method are discussed in the following subsections.

A. Feature Extraction

Given an expression video clip, the subject's face is detected and fitted using discriminative response map fitting (DRMF), which has been shown to outperform other face fitting methods [45]. Also, DRMF is computationally efficient and hence can be used to process a large number of videos. Next, the face is cropped using the landmark points to eliminate background information, resulting in an aligned video where dense trajectories are computed as explained below.

First, feature points are densely sampled on a grid spaced by $W = 5$ pixels [31] in different spatial scales. There are at most 8 spatial scales increasing by a factor of $1/\sqrt{2}$ and the actual number of scales used depends on the resolution of the video. Feature points are tracked on each spatial scale separately using dense optical flow. For each frame I_t, its dense optical flow field $\omega_t = (u_t, v_t)$ is computed w.r.t. the next frame I_{t+1}, where u_t and v_t are the horizontal and vertical components of the optical flow. Given a point $P_t = (x_t, y_t)$ in frame I_t, its tracked position in frame I_{t+1} is smoothed by applying a median filter on ω_t:

$$P_{t+1} = (x_{t+1}, y_{t+1}) = (x_t, y_t) + (\mathbf{K} * \omega_t)|_{(x_t, y_t)}, \quad (1)$$

where $\mathbf{K}$ is the median filtering kernel of size 3×3 pixels. Once the dense optical flow field is computed, points can

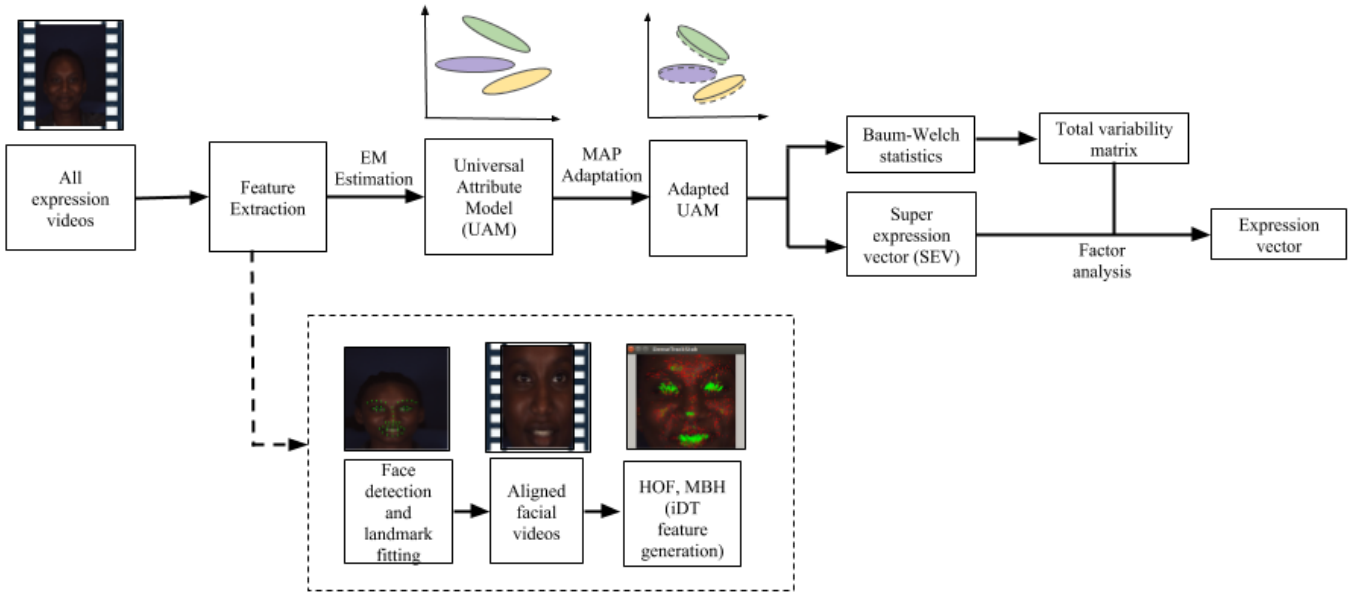


Fig. 1: Block diagram of the proposed expression-vector extraction using universal attribute modelling (best viewed in colour).

be tracked very densely without additional cost. The points in subsequent frames are concatenated to form trajectories $(P_t, P_{t+1}, P_{t+2}, ...)$. However, if no tracked point is found in a $W \times W$ neighbourhood, a new point is sampled and added to the tracking process so that a dense coverage of trajectories is ensured. As trajectories tend to drift from their initial locations during the tracking process, we limit their length to $L = 15$ frames to overcome this problem [46]. Also, trajectories with sudden large displacements are most likely to be erroneous. Hence, if the consecutive frame displacement > 70 % of the overall displacement of the trajectory, the trajectory is removed.

Local descriptors like HOG3D, 3DSIFT, and LBP-TOP are usually computed in a 3D video volume around interest points, which ignore the fundamental dynamic structures in the video. Hence, the HOF and MBH descriptors are calculated here within a space-time volume of size $N \times N$ pixels, where $N = 32$ and L frames [46] aligned with a trajectory as shown in Figure 2. To obtain local structural information, this volume is subdivided into cells of size $n_h \times n_w \times n_t$, where $n_h = 2$, $n_w = 2$, and $n_t = 3$ are height, width, and temporal segment lengths. We compute a descriptor (HOF and MBH) in each cell of the space-time volume.

For HOF, orientations are quantized into a total of 9 bins and are normalized by its L_2 norm. The final descriptor size is 108 for HOF. The MBH descriptor encodes the gradient of the optical flow, which results in the removal of locally constant camera motion and the retention of information about changes in the flow field (i.e., motion boundaries). The MBH descriptor separates optical flow $\omega = (u, v)$ into its horizontal and vertical components. Spatial derivatives are computed for each of them, and orientation information is quantized into histograms, and the magnitude is used for weighting. We obtain an 8-bin histogram for each component (i.e., MBHx and MBHy). Both histogram vectors are normalized separately with their L_2 norm. The dimension obtained for both MBHx and MBHy is 96 (i.e., $2 \times 2 \times 3 \times 8$). After computing HOF and MBH descriptors, a universal attribute model (UAM) is constructed for each descriptor as described in the next subsection.

B. Universal Attribute Model (UAM)

We can consider each expression clip to be a sample function, which realizes the random process generating the expression. To compare the similarity of two expression clips, we need to match the sample functions. Such a match can be based only on the parameters of the probability density function (*pdf*) that describes the random process generating the expression. If we assume that the underlying *pdf* can be estimated using a GMM, then the number of mixtures must be sufficiently large to accommodate the intra-expression variances encountered in spontaneous expressions. Unfortunately, a single clip does not have enough data points to estimate the *pdf* of the expression. Hence, we propose to train a universal GMM using the clips of all the expressions. We call this model the universal attribute model (UAM), which has a large number of mixtures for modelling the attributes of different expressions. However, it is observed that as attributes are shared across expressions, even a modest number of mixtures is enough for achieving good representation.

The universal attribute model (UAM) can be represented as

$$p(\mathbf{x}_l) = \sum_{c=1}^{C} w_c \mathcal{N}(\mathbf{x}_l | \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c), \tag{2}$$

where the mixture weights w_c satisfy the constraint $\sum_{c=1}^{C} w_c = 1$ and $\boldsymbol{\mu}_c, \boldsymbol{\sigma}_c$ are the mean and covariance for mixture c of the UAM, respectively. A feature $\mathbf{x}_l$ is part of a clip $\mathbf{x}$ represented as a set of feature vectors $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L$. This feature can be either an HOF or an MBH descriptor and

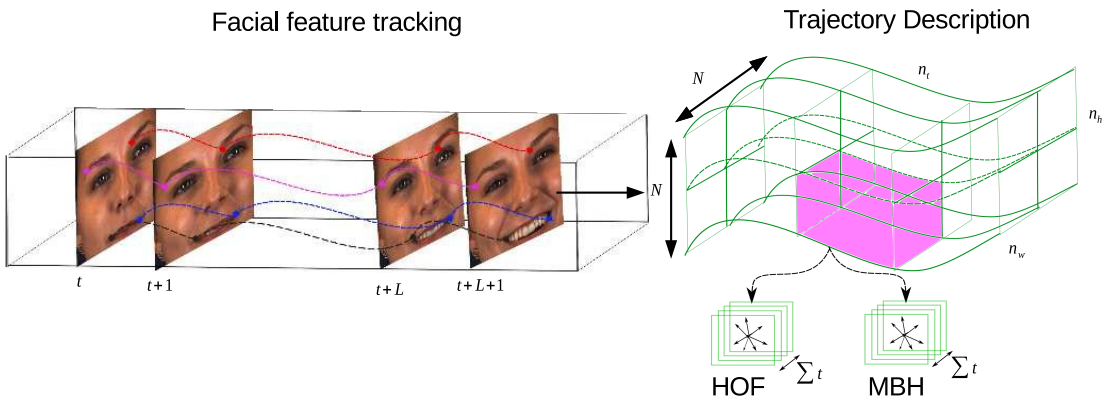


Fig. 2: HOF and MBH extraction from trajectories. Adapted from [46]. Best viewed in colour.

we train a separate UAM for each descriptor during evaluation using standard EM estimation.

We hypothesize that after training the UAM, each Gaussian component in the UAM captures an attribute. This attribute can be specific to a particular expression or may be present in multiple expressions. As the goal is still to find the *pdf* of the expression that generates a clip, we need to adapt the UAM parameters using the data in the clip [47], [48].

C. Super expression-vector (SEV) representation

The UAM parameters are adapted for each mixture component given L feature vectors of a clip $\mathbf{x}$. The probabilistic alignment of these feature vectors into each of the C mixture components of the UAM is calculated as a posterior $p(c|\mathbf{x}_l)$ given by

$$p(c|\mathbf{x}_l) = \frac{w_c p(\mathbf{x}_l|c)}{\sum_{c=1}^{C} w_c p(\mathbf{x}_l|c)}, \tag{3}$$

where $\mathbf{x}_l$ is a $d \times 1$ feature vector and $p(\mathbf{x}_l|c)$ is the likelihood of a feature $\mathbf{x}_l$ arriving from a mixture c.

The posterior probability is then used to calculate the zeroth and first order Baum-Welch statistics for a clip $\mathbf{x}$ given as $n_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)$ and $\mathbf{F}_c(\mathbf{x}) = (\sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{x}_l)/n_c(\mathbf{x})$, respectively. The MAP adapted parameters of a clip-specific model can be obtained as a convex combination of the UAM and the clip-specific statistics. For every mixture c of the UAM, the adapted weights and means are calculated as

$$\hat{w}_c = \alpha n_c(\mathbf{x})/L + (1-\alpha)w_c \tag{4a}$$

and

$$\hat{\boldsymbol{\mu}}_c = \alpha \mathbf{F}_c(\mathbf{x}) + (1-\alpha)\boldsymbol{\mu}_c. \tag{4b}$$

The covariance is not modified as there is not enough data in one clip to update the entire covariance matrix of the UAM. The adapted means for each mixture are then concatenated to compute a $(Cd \times 1)$-dimensional SEV for each clip $\mathbf{s}(\mathbf{x}) = [\hat{\boldsymbol{\mu}}_1 \hat{\boldsymbol{\mu}}_2 \cdots \hat{\boldsymbol{\mu}}_C]^t$. Obtaining a fixed-dimensional representation like the super expression-vector normalizes the effect of varying length clips but results in a high-dimensional representation. This representation though contains many of the attributes that do not contribute to the clip and hence are not changed from the original UAM. Since each clip contains only a few of the total UAM mixtures (attributes), only those means are modified. Hence, the SEV is intrinsically low-dimensional, and by using a suitable decomposition, we can extract such a representation, which we refer to as an expression-vector.

D. Expression-vector representation

In order to arrive at a low-dimensional representation, the super-expression vector $\mathbf{s}$ is decomposed as

$$\mathbf{s} = \mathbf{m} + \mathbf{T}\mathbf{w}, \tag{5}$$

where $\mathbf{m}$ is the supervector that is actor and viewpoint independent (can be assumed to be the unadapted UAM supervector), $\mathbf{T}$ is a low-rank rectangular matrix known as the total variability matrix of size $Cf \times r$, and an r-dimensional random vector $\mathbf{w}$ whose prior distribution is assumed to be a standard Gaussian $\mathcal{N}(\mathbf{0}, \mathbf{I})$ [49]. We refer to this random vector as an *expression-vector*, which is a hidden variable and is defined by its posterior distribution $P(\mathbf{w}|\mathbf{x})$ after observing a clip $\mathbf{x}$ as

$$P(\mathbf{w}|\mathbf{x}) \propto P(\mathbf{x}|\mathbf{w})\mathcal{N}(\mathbf{0}, \mathbf{I}) \tag{6}$$

$$\propto \exp\left(-\frac{1}{2}(\mathbf{w} - \mathbf{L}(\mathbf{x}))^t \mathbf{M}(\mathbf{x})(\mathbf{w} - \mathbf{L}(\mathbf{x}))\right) \tag{7}$$

where $\boldsymbol{\Sigma}$ is a diagonal covariance matrix of dimension $Cd \times Cd$ and it models the residual variability not captured by the total variability matrix $\mathbf{T}$. The matrix $\mathbf{L}(\mathbf{x}) = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x})$, where $\tilde{\mathbf{s}}(\mathbf{x})$ is the centered supervector, which appears because the posterior distribution of $\mathbf{w}$ is conditioned on the Baum-Welch statistics of the clip centered around the means of the UAM. The first order Baum-Welch statistics centered around the UAM mean can be obtained as $\tilde{\mathbf{F}}_c(\mathbf{x}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)$. We can now express $\tilde{\mathbf{s}}(\mathbf{x})$ as the concatenated first-order statistics $\tilde{\mathbf{s}}(\mathbf{x}) = [\tilde{\mathbf{F}}_1(\mathbf{x})\tilde{\mathbf{F}}_2(\mathbf{x}) \cdots \tilde{\mathbf{F}}_C(\mathbf{x})]^t$. Also, the matrix $\mathbf{M}(\mathbf{x}) = \mathbf{I} + \mathbf{T}^t\boldsymbol{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{T}$, where $\mathbf{N}(\mathbf{x})$ is a diagonal matrix of dimension $Cd \times Cd$ whose diagonal blocks are $n_c(\mathbf{x})\mathbf{I}$, for $c = 1, ..., C$ and $\mathbf{I}$ is the identity matrix of dimension $d \times d$.

From Equation 7, the mean and covariance matrix of the posterior distribution are given by

$$E[\mathbf{w}(\mathbf{x})] = \mathbf{M}^{-1}(\mathbf{x})\mathbf{T}^t\boldsymbol{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}) \tag{8a}$$

and

$$\mathrm{Cov}(\mathbf{w}(\mathbf{x}), \mathbf{w}(\mathbf{x})) = \mathbf{M}^{-1}(\mathbf{x}), \tag{8b}$$

respectively. Using EM algorithm [50], we iteratively estimate the posterior mean and covariance in the E-step and use the same to update $\mathbf{T}$ and $\mathbf{\Sigma}$ in the M-step.

In the first E-step of the estimation, $\mathbf{m}$ and $\mathbf{\Sigma}$ are initialized with the UAM mean and covariance, respectively. For the total variability matrix $\mathbf{T}$, a desired rank r is chosen, and the matrix is initialized randomly. Then $E[\mathbf{w}(\mathbf{x})]$ and $\mathrm{Cov}(\mathbf{w}(\mathbf{x}), \mathbf{w}(\mathbf{x}))$ are calculated according to Equations 8a & 8b.

In the M-step, the matrix $\mathbf{T}$ is calculated as the solution of

$$\sum_{\mathbf{x}} \mathbf{N}(\mathbf{x})\mathbf{T}E[\mathbf{w}(\mathbf{x})\mathbf{w}^t(\mathbf{x})] = \sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})], \tag{9}$$

which results in a system of r linear equations. The right hand side of Equation 9 contains $\tilde{\mathbf{s}}(\mathbf{x})$, accounting for the number of features in the clip. As $\mathbf{T}$ is the same for all the clips, the left hand side is also weighed by $\mathbf{N}(\mathbf{x})$ to account for the number of features in the clip.

For each $c = 1, \cdots, C$, the residual matrix $\mathbf{\Sigma}$ is estimated mixture by mixture as

$$\mathbf{\Sigma}_c = \frac{1}{n_c(\mathbf{x})} \left(\sum_{\mathbf{x}} \tilde{S}_c(\mathbf{x}) - \mathbf{M}_c \right), \tag{10}$$

where $\mathbf{M}_c$ denotes the c^{th} diagonal block of the $Cd \times Cd$ matrix $\frac{1}{2}\sum_{\mathbf{x}} \tilde{\mathbf{s}}(\mathbf{x})E[\mathbf{w}^t(\mathbf{x})]T^t + TE[\mathbf{w}(\mathbf{x})]\tilde{\mathbf{s}}^t(\mathbf{x})$. And, $\tilde{\mathbf{S}}_c(\mathbf{x})$ is the second-order Baum-Welch statistics of the clip centered on the means of the UAM calculated as

$$\tilde{\mathbf{S}}_c(\mathbf{x}) = diag\left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)(\mathbf{x}_l - \boldsymbol{\mu}_c)(\mathbf{x}_l - \boldsymbol{\mu}_c)^t \right). \tag{11}$$

After the final M-step i.e. estimation of $\mathbf{T}$ and $\mathbf{\Sigma}$ matrices, the expression-vector for a given clip can be represented using the mean of its posterior distribution as

$$\mathbf{w}(\mathbf{x}) = (\mathbf{I} + \mathbf{T}^t\mathbf{\Sigma}^{-1}\mathbf{N}(\mathbf{x})\mathbf{T})^{-1}\mathbf{T}^t\mathbf{\Sigma}^{-1}\tilde{\mathbf{s}}(\mathbf{x}). \tag{12}$$

This process of obtaining the expression-vector is known as factor analysis [50]. The $\mathbf{T}$-matrix contains the eigenvectors of the largest r eigenvalues of the total variability covariance matrix [49]. We hypothesize that these large eigenvalues arrive from the Gaussian mixture(s), which model the attributes in the clip. The original SEV can now be projected onto a r-dimensional expression-vector based on T. In the next subsection, we explore various scoring mechanisms for expression-vectors like cosine scoring, linear discriminant analysis, and probabilistic linear discriminant analysis.

IV. Expression-vector scoring

The main objective of this work is to compare expression clips based on the underlying *pdf* of the expressions. The *pdf* was estimated using a GMM (UAM) where each mixture is assumed to learn an attribute. Using an adapted UAM and factor analysis, we arrived at a representation called expression vector, which contains only useful attributes for that expression clip. Once such a normalised representation is obtained (because of $\mathbf{N}(\mathbf{x})$ in Equation 12), we can directly compare the expression-vectors of any two clips using cosine scoring.

Cosine scoring. For cosine scoring, the distance between a pair of expression-vectors is given as

$$k(\mathbf{w}_1, \mathbf{w}_2) = \frac{\mathbf{w}_1^t\mathbf{w}_2}{\sqrt{\mathbf{w}_1^t\mathbf{w}_1}\sqrt{\mathbf{w}_2^t\mathbf{w}_2}}. \tag{13}$$

Linear Discriminant Analysis (LDA). While cosine scoring requires no class labels for evaluation, it cannot address intra-expression variability. Such variability arises when recording the same expression from different camera view that can reveal or hide certain expression attributes. We hypothesize that using class information in multi-class linear discriminant analysis can reduce the effect of intra-class variability. Hence, the coefficients of the projection matrix $\hat{\mathbf{A}}$ are chosen so as to maximize the ratio of the between-class scattering $\mathbf{S}_b$ to the within-class $\mathbf{S}_w$ scattering

$$\hat{\mathbf{A}} = \underset{\mathbf{A}}{\mathrm{argmax}} \frac{\mathbf{A}^t\mathbf{S}_b\mathbf{A}}{\mathbf{A}^t\mathbf{S}_w\mathbf{A}}. \tag{14}$$

The solution to the above maximization problem is given by the following generalized eigenvalue problem:

$$\mathbf{S}_b\mathbf{A} = \lambda\mathbf{S}_w\mathbf{A}. \tag{15}$$

Generally, at most $(m-1)$ generalized eigenvectors of $\mathbf{A}$ are useful to discriminate among m expressions.

Probabilistic Linear Discriminant Analysis (PLDA). Even though LDA solves the intra-expression variability issue, the number of dimensions available for projection is always limited by the number of classes. In PLDA, there is no dimensionality constraint and it has been shown to produce better results than LDA for tasks like face recognition [51], [52] and object recognition [53]. Hence, we propose to use PLDA based expression-vector scoring, which is derived with a two-covariance model similar to LDA. The two-covariance model is known as a generative linear-Gaussian model, where latent vectors $\mathbf{y}$ representing expressions are assumed to be distributed according to the prior distribution

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{y}|\boldsymbol{\mu}, \mathbf{S}_b). \tag{16}$$

For a given expression represented by a latent vector $\hat{\mathbf{y}}$, the distribution of expression-vector $\mathbf{w}$ is assumed to be

$$p(\mathbf{y}) = \mathcal{N}(\mathbf{w}|\mathbf{y}, \mathbf{S}_w). \tag{17}$$

The maximum-likelihood estimates of the model parameters, $\boldsymbol{\mu}, \mathbf{S}_b$, and $\mathbf{S}_w$, can be obtained using EM algorithm. Now, in case of LDA, the projection of all the vectors using the projection matrix $\mathbf{w}$ would be followed by cosine scoring. In PLDA, the projection matrix $\mathbf{w}$ is not obtained explicitly and scoring is done using every pair of expression-vectors ($\mathbf{w}_1$,$\mathbf{w}_2$) using the following two hypotheses:

- *Null hypothesis* $\mathscr{H}_s$: A single latent vector $\hat{\mathbf{y}}$ representing the expression is generated from the prior $p(\mathbf{y})$, for which both $\mathbf{w}_1$ and $\mathbf{w}_2$ are generated from $p(\mathbf{w}|\hat{\mathbf{y}})$.
- *Alternative hypothesis* $\mathscr{H}_d$: Two latent vectors representing two different expressions are independently generated from $p(\mathbf{y})$. For each latent vector, either $\mathbf{w}_1$ or $\mathbf{w}_2$ is

generated.

The score can now be calculated as a log-likelihood ratio between the two hypotheses $\mathscr{H}_s$ and $\mathscr{H}_d$ as

$$k(\mathbf{w}_1, \mathbf{w}_2) = log \frac{p(\mathbf{w}_1, \mathbf{w}_2|\mathscr{H}_s)}{p(\mathbf{w}_1, \mathbf{w}_2|\mathscr{H}_d)} \quad (18)$$

$$= log \frac{\int p(\mathbf{w}_1|\mathbf{y})p(\mathbf{w}_2|\mathbf{y})p(\mathbf{y})d\mathbf{y}}{p(\mathbf{w}_1)p(\mathbf{w}_2)}, \quad (19)$$

where in the numerator, we integrate over the distribution of expression-vectors to determine the likelihood of producing both expression-vectors having the same expression. For the hypothesis that the expression-vector belongs to separate expressions, the product of the marginals $p(\mathbf{w}_1)$ and $p(\mathbf{w}_2)$ is used.

V. Experiments

A. Datasets and protocols

To show the applicability of the proposed method on different recording conditions, we chose two datasets - BP4D, which is recorded in a controlled setup and AFEW, which is compiled from videos shot in an unconstrained environment.

BP4D- The Binghamton Pittsburgh 4D (BP4D) spontaneous expression dataset [19] is recorded in a controlled environment where the subjects respond naturally to different video clips being played in front of them. The dataset consists of 41 subjects, captured from 9 different views with 8 distinct tasks (where each task is meant to elicit a particular expression) resulting in a total of 2952 videos. As per the protocol followed in [19], 39 subjects were used for training, and 2 (1 male) were used for testing. We focused on recognizing five emotions, namely, anger, fear, happiness, sadness, and surprise. The main reason for selecting BP4D database is because it is the first dynamic and spontaneous facial expression database, which is not deliberately posed and contains subjects across different age groups and different countries.

AFEW- Acted Facial Expression in the Wild (AFEW) [20] is a dynamic video dataset created from Hollywood movies containing 723 training videos and 383 validation videos. The videos are completely unconstrained and provide a fair idea of real world conditions in which expressions can be captured.

In all the results reported below, the expression-vector is 200 dimensional, which was determined empirically as the best value. For PLDA, the best projection dimension was found to be 150, and in case of LDA, the number of dimensions is 4, which is one less than the total number of expressions.

B. Analysis of expression-vector on dataset specific UAM

In Table I and Figure 3, the classification performance of expression-vector is presented on both BP4D and AFEW datasets. There are 4 UAMs which are trained for this purpose, 2 trained each on BP4D and AFEW datasets using both HOF and MBH descriptors. The best performance of expression-vector for BP4D is obtained using HOF descriptor for 256 UAM mixtures and SVM classifier. On the other hand, for the AFEW dataset, 64 UAM mixtures trained with either HOF or MBH descriptor and PLDA classifier gives the best classification accuracy. Hence, it can be observed that a UAM with a smaller number of mixtures is suitable for AFEW while a larger number of mixtures benefits BP4D. This leads us to conjecture that as the number of views in AFEW is less than BP4D, it gives rise to fewer attributes. Also, it can be observed that increasing the number of mixtures does not cause much difference in classification for AFEW dataset, which supports our hypothesis.

C. Cross-dataset evaluation

In nearly all expression recognition settings, most of the labelled training examples are obtained from constrained environments. However, in real-world situations, the instances to be classified are mostly obtained from unconstrained environments. So, in this work, we attempted to recognize expressions in an unconstrained dataset, i.e. AFEW using an attribute model trained on a constrained dataset, i.e. BP4D. We used the UAM and total variability matrix (Equation 5) trained on the BP4D dataset to form expression-vectors for the unconstrained videos in the AFEW dataset. The classification accuracy on AFEW dataset is presented in Table II and Figure 4. It is interesting to note that the best classification performance surpasses what was obtained for the UAM trained on AFEW. This shows that BP4D has more diverse attributes than AFEW, which results in a more discriminative expression-vector. Hence, we hypothesize that both UAM and T-matrix trained on enough diverse examples generalize well for unseen data. Also, in cases where unconstrained data is not sufficient to train a UAM, one that is trained on constrained data can be effectively used.

D. Analysis of expression-vector on combined UAM

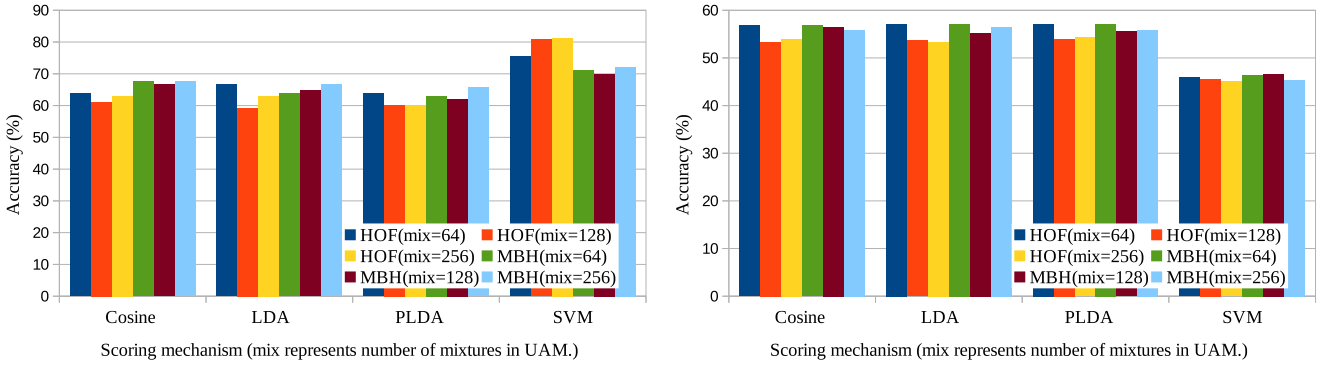
As BP4D captures spontaneous expressions in a controlled environment and AFEW captures them in an unconstrained environment, we explore whether they may contain complimentary attributes which can benefit the expression-vector representation. The combined UAM and total variability matrix are trained using the training data of both BP4D and AFEW datasets. In Table III, the classification performance of expression-vector extracted from a combined UAM is presented. It can be observed that there is no improvement in classification performance in the case of BP4D or AFEW when compared to the individual UAMs. So, it can be concluded that there are no complimentary attributes that are modelled using the combined UAM to arrive at a better expression-vector.

E. Comparison with state-of-the-art approaches

The comparison of the proposed method with state-of-the-art techniques is presented for BP4D and AFEW datasets in Table IV and V, respectively. The existing approaches depend on an ensemble of low-level features [35], [54], annotated facial action units [55], multi-modal features (face, facial actions, and audio) [55], [56], and polynomial fitting of spatio-temporal volumes (for view invariance) [19] to achieve state-of-the-art performance. However, such representations often suffer from noise and spatio-temporal sensitivity, which lead

TABLE I: Expression-vector classification performance (in %) for BP4D and AFEW with dataset specific UAM.

Scoring mechanism	Number of UAM mixtures											
	BP4D						AFEW					
	HOF			MBH			HOF			MBH		
	64	128	256	64	128	256	64	128	256	64	128	256
Cosine	63.89	61.11	62.96	67.59	66.67	67.59	56.79	53.31	53.94	56.79	56.47	55.84
LDA	66.67	59.26	62.96	63.88	64.81	66.67	**57.10**	53.63	53.32	**57.10**	55.21	56.47
PLDA	63.89	60.18	60.19	62.97	62.04	65.75	**57.10**	53.94	54.26	**57.10**	55.52	55.84
SVM	75.40	80.80	**81.30**	71.00	69.90	72.20	45.90	45.60	45.00	46.40	46.60	45.40



(a) BP4D (b) AFEW

Fig. 3: Classification performance of expression-vector with dataset specific UAM. Best viewed in colour.

TABLE II: Expression-vector on-line classification performance (in %) on AFEW. UAM learned using training data from BP4D.

Scoring mechanism	Number of UAM mixtures					
	HOF			MBH		
	64	128	256	64	128	256
Cosine	54.26	52.68	54.57	52.05	51.42	50.16
LDA	54.57	52.37	54.26	50.47	50.79	50.16
PLDA	54.26	53	53.94	50.79	52.37	50.47
SVM	72.6	72.8	**74.1**	65.2	67.40	68.60

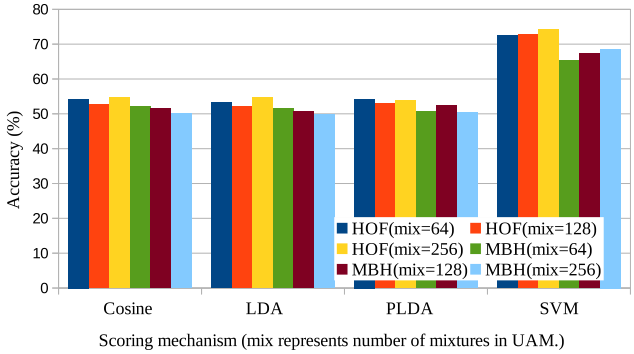


Fig. 4: Classification performance of expression-vector on AFEW with UAM and T-matrix learned using training data from BP4D. Best viewed in colour.

to poor performance, especially on an unconstrained dataset like AFEW as compared to the proposed expression-vectors as can be seen in Table V. It should also be noted that expression-vectors achieve view invariance without the need for polynomial fitting as in [19]. This is because the UAM model trained in the proposed approach allows for a view normalized expression representation.

The closest approach to the proposed expression-vector is an ensemble of mid-level representation called expressionlet [30] which was discussed in Section I. However, expression-vectors outperform the expressionlet ensemble on unconstrained (AFEW) and spontaneous (BP4D) datasets. This can be attributed to the fact that the formation of expression-vectors uses adapted means of the UAM whereas expressionlets are formed using the means of chosen local features per mixture of the universal model. Further, we hypothesize that the factor analysis can preserve important attributes better than discriminative learning used for forming the ensemble of expressionlets. From Tables IV and V, it can be observed that expression-vectors combined with an unsupervised scoring technique like cosine scoring, perform better on constrained and unconstrained datasets as compared to supervised classification of expressionlet ensemble.

F. Expression-wise comparison

The confusion matrices for the best performance on BP4D and AFEW are presented in Figure 5 (a) and 5 (b), respectively. It can be seen that the classification performance for all the expressions is close to the overall classification accuracy, which shows that UAM captures the attributes of all expressions equally well. In particular, we can observe misclassification of *happy* clips as *angry* and vice-versa for both the datasets. As shown in the Figure 6 (a), there is an overlap in the

TABLE III: Expression-vector classification performance (in %) for BP4D and AFEW with combined UAM.

Scoring mechanism	Number of UAM mixtures											
	BP4D						AFEW					
	HOF			MBH			HOF			MBH		
	64	128	256	64	128	256	64	128	256	64	128	256
Cosine	71.30	62.04	60.19	67.59	64.81	62.03	52.36	53.94	50.47	**55.84**	53.94	54.26
LDA	60.18	63.89	62.037	65.74	64.81	66.67	52.36	51.41	51.10	53.31	51.73	55.20
PLDA	62.03	73.15	74.07	59.26	63.89	62.04	48.26	50.47	48.58	48.89	48.26	49.21
SVM	74.20	75.50	**81.00**	74.20	73.70	76.90	44.50	44.80	46.40	43.30	43.80	43.60

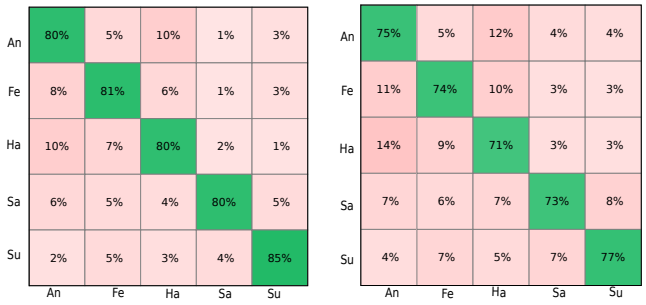
TABLE IV: State-of-the-art on BP4D dataset

Methods	Accuracy (%)
3DCNN* [57]	46.5
HOG + HMM [58]	72.2
Nebula + LDA [19]	76.1
Rotation reversal HOG + conditional random forests [59]	76.8
Expression-vector +cosine	71.3
Expression-vector +SVM	**81.3**

* our evaluation using C3D features

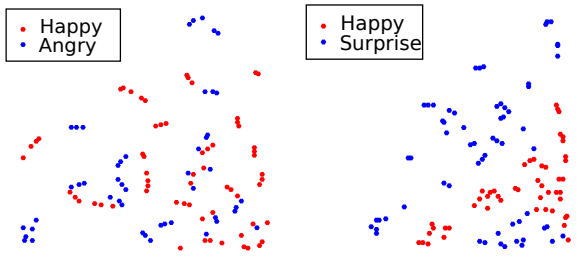
TABLE V: State-of-the-art on AFEW dataset

Methods	Accuracy (%)
3DCNN [60]	31.3
Expressionlet [61]	31.7
Multi-modal CNN + DNN [56]	49.5
Feature Ensemble + Partial Least Squares [35]	52.3
AU + Audio + SVM [55]	53.8
HOG + Boosted cascade [62]	56.8
Expression-vector +cosine	56.8
Expression-vector +SVM	**74.1**



(a) BP4D dataset (b) AFEW dataset

Fig. 5: Confusion matrix for expression-vector with best model (HOF features on 256 components, classified using SVM). Best viewed in colour.



(a) (b)

Fig. 6: t-sne plot for expression-vectors of (a) *happy* vs. *angry* and (b) *happy* vs. *surprise* of BP4D dataset.

extracted expression-vectors for *happy* and *angry* which results in this confusion. However, in the case of Figure 6 (b), there is almost no overlap in the expression-vectors for *happiness* and *surprise*, which reduces the confusion between these two classes.

VI. Conclusion

In this work, we proposed a novel actor-independent, low-dimensional, and high-level representation for spontaneous expression recognition termed expression-vector. With no pre-processing and manual annotation, feature descriptors were obtained, and an entirely unsupervised learning mechanism was shown for obtaining expression-vectors. On BP4D and AFEW datasets, expression-vectors demonstrated better performance than state-of-the-art approaches, which shows the ability of the proposed approach in capturing discriminative information without supervision. Also, we have shown that a UAM trained on a constrained dataset can be used to recognize unconstrained expression clips effectively. In future, we would like to investigate the performance of expression-vector for localization of expressions in temporally untrimmed clips.

References

[1] D. Matsumoto, J. Leroux, C. Wilson-cohn, J. Raroque, K. Kooken, P. Ekman, N. Yrizarry, S. Loewinger, H. Uchida, A. Yee, L. Amo, A. Goh, D. Matsumoto, J. Leroux, C. Wilson-cohn, J. Raroque, K. Kooken, S. Loewinger, H. Uchida, A. Yee, L. Amo, A. Goh, W. T. C. Kim, and S. Paul, "A new test to measure emotion recognition ability: Matsumoto and ekmans japanese and caucasian brief affect recognition test (jacbart," *Journal of Nonverbal Behavior*, pp. 179–209, 2000. 1

[2] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Proceedings of the 7th International Conference on Automatic Face and Gesture Recognition*, ser. FGR '06. Washington, DC, USA: IEEE Computer Society, 2006, pp. 211–216. [Online]. Available: http://dx.doi.org/10.1109/FGR.2006.6 1

[3] M. S. Bartlett, G. Littlewort, M. G. Frank, C. Lainscsek, I. R. Fasel, and J. R. Movellan, "Automatic recognition of facial actions in spontaneous expressions." *Journal of multimedia*, vol. 1, no. 6, pp. 22–35, 2006. 1

[4] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," *Image Vision Comput.*, vol. 27, no. 12, pp. 1797–1803, Nov. 2009. [Online]. Available: http://dx.doi.org/10.1016/j.imavis.2008.12.010 1

[5] G. McKeown, M. F. Valstar, R. Cowie, M. Pantic, and M. Schröder, "The

semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent," *IEEE Transactions on Affective Computing*, vol. 3, pp. 5–17, 2012. 1

[6] N. Perveen, D. Singh, and C. K. Mohan, "Spontaneous facial expression recognition: A part based approach," in *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, Dec 2016, pp. 819–824. 1

[7] N. Perveen, S. Gupta, and K. Verma, "Facial expression recognition using facial characteristic points and gini index," in *2012 Students Conference on Engineering and Systems*, March 2012, pp. 1–6. 1

[8] H. Dibeklioglu, A. A. Salah, and T. Gevers, "Recognition of genuine smiles," *IEEE Transactions on Multimedia*, vol. 17, no. 3, pp. 279–294, March 2015. 1

[9] I. Ofodile, K. Kulkarni, C. A. Corneanu, S. Escalera, X. Baro, S. Hyniewska, J. Allik, and G. Anbarjafari, "Automatic recognition of deceptive facial expressions of emotion," *arXiv preprint arXiv:1707.04061*, 2017. 1

[10] C. A. Corneanu, M. O. Simn, J. F. Cohn, and S. E. Guerrero, "Survey on rgb, 3d, thermal, and multimodal approaches for facial expression recognition: History, trends, and affect-related applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1548–1568, Aug 2016. 1

[11] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A survey of affect recognition methods: Audio, visual, and spontaneous expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 31, no. 1, pp. 39–58, 2009. 1

[12] H. Nguyen, K. Kotani, F. Chen, and B. Le, "A thermal facial emotion database and its analysis," in *Pacific-Rim Symposium on Image and Video Technology*. Springer, 2013, pp. 397–408. 1

[13] P. Ekman, J. C. Hager, and W. V. Friesen, "The symmetry of emotional and deliberate facial actions," *Psychophysiology*, vol. 18, no. 2, pp. 101–106, 1981. 1

[14] P. Ekman and W. V. Friesen, "Felt, false, and miserable smiles," *Journal of Nonverbal Behavior*, vol. 6, no. 4, pp. 238–252, Jun 1982. [Online]. Available: https://doi.org/10.1007/BF00987191 1

[15] J. F. COHN and K. L. SCHMIDT, "The timing of facial motion in posed and spontaneous smiles," *International Journal of Wavelets, Multiresolution and Information Processing*, vol. 02, no. 02, pp. 121–132, 2004. [Online]. Available: http://www.worldscientific.com/doi/abs/10.1142/S021969130400041X 1

[16] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (ck+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, June 2010, pp. 94–101. 1

[17] L. Yin, X. Wei, Y. Sun, J. Wang, and M. J. Rosato, "A 3d facial expression database for facial behavior research," in *Automatic face and gesture recognition, 2006. FGR 2006. 7th international conference on*. IEEE, 2006, pp. 211–216. 1

[18] L. Yin, X. Chen, Y. Sun, T. Worm, and M. Reale, "A high-resolution 3d dynamic facial expression database," in *2008 8th IEEE International Conference on Automatic Face Gesture Recognition*, Sept 2008, pp. 1–6. 1

[19] X. Zhang, L. Yin, J. F. Cohn, S. Canavan, M. Reale, A. Horowitz, P. Liu, and J. M. Girard, "Bp4d-spontaneous: a high-resolution spontaneous 3d dynamic facial expression database," *Image and Vision Computing*, vol. 32, no. 10, pp. 692–706, 2014. 1, 6, 7, 8

[20] A. Dhall, R. Goecke, S. Lucey, and T. Gedeon, "Collecting large, richly annotated facial-expression databases from movies," *IEEE MultiMedia*, vol. 19, no. 3, pp. 34–41, July 2012. 1, 6

[21] D. McDuff, R. el Kaliouby, T. Senechal, M. Amr, J. F. Cohn, and R. Picard, "Affectiva-mit facial expression dataset (am-fed): Naturalistic and spontaneous facial expressions collected in-the-wild," in *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops*, June 2013, pp. 881–888. 1

[22] R. A. Calvo and S. D'Mello, "Affect detection: An interdisciplinary review of models, methods, and their applications," *IEEE Transactions on Affective Computing*, vol. 1, no. 1, pp. 18–37, Jan 2010. 1

[23] Y. Li, J. Chen, Y. Zhao, and Q. Ji, "Data-free prior model for facial action unit recognition," *IEEE Transactions on Affective Computing*, vol. 4, no. 2, pp. 127–141, April 2013. 1

[24] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, June 2007. 1

[25] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8. 1

[26] A. Klaser, M. Marszalek, and C. Schmid, "A Spatio-Temporal Descriptor Based on 3D-Gradients," in *BMVC 2008 - 19th British Machine Vision Conference*, M. Everingham, C. Needham, and R. Fraile, Eds. Leeds, United Kingdom: British Machine Vision Association, Sep. 2008, pp. 275:1–10. [Online]. Available: https://hal.inria.fr/inria-00514853 1

[27] P. Scovanner, S. Ali, and M. Shah, "A 3-dimensional sift descriptor and its application to action recognition," in *Proceedings of the 15th ACM International Conference on Multimedia*, ser. MM '07. New York, NY, USA: ACM, 2007, pp. 357–360. [Online]. Available: http://doi.acm.org/10.1145/1291233.1291311 1

[28] P. Ekman, *Emotions Revealed: Recognizing Faces and Feelings to Improve Communication and Emotional Life*. Henry Holt and Company, 2003. [Online]. Available: https://books.google.co.in/books?id=JJ9xHXo7hPgC 1

[29] Y. J. Liu, J. K. Zhang, W. J. Yan, S. J. Wang, G. Zhao, and X. Fu, "A main directional mean optical flow feature for spontaneous micro-expression recognition," *IEEE Transactions on Affective Computing*, vol. 7, no. 4, pp. 299–310, Oct 2016. 1

[30] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets via universal manifold model for dynamic facial expression recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 12, pp. 5920–5932, Dec 2016. 1, 7

[31] H. Wang and C. Schmid, "Action recognition with improved trajectories," in *Proceedings of the IEEE International Conference on Computer Vision*, 2013, pp. 3551–3558. 2

[32] D. Li, H. Zhou, and K. M. Lam, "High-resolution face verification using pore-scale facial features," *IEEE Transactions on Image Processing*, vol. 24, no. 8, pp. 2317–2327, Aug 2015. 2

[33] S. M. Lajevardi and H. R. Wu, "Facial expression recognition in perceptual color space," *IEEE Transactions on Image Processing*, vol. 21, no. 8, pp. 3721–3733, Aug 2012. 2

[34] F. Tsalakanidou and S. Malassiotis, "Real-time facial feature tracking from 2d and 3d video streams," in *2010 3DTV-Conference: The True Vision - Capture, Transmission and Display of 3D Video*, June 2010, pp. 1–4. 2

[35] H. Kaya, F. Gürpinar, S. Afshar, and A. A. Salah, "Contrasting and combining least squares based learners for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 459–466. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830588 2, 6, 8

[36] A. Dhall, A. Asthana, R. Goecke, and T. Gedeon, "Emotion recognition using phog and lpq features," in *Face and Gesture 2011*, March 2011, pp. 878–883. 2

[37] X. Feng, M. Pietikäinen, and A. Hadid, "Facial expression recognition based on local binary patterns," *Pattern Recognition and Image Analysis*, vol. 17, no. 4, pp. 592–598, 2007. 2

[38] M. Kchele, M. Schels, P. Thiam, and F. Schwenker, "Fusion mappings for multimodal affect recognition," in *2015 IEEE Symposium Series on Computational Intelligence*, Dec 2015, pp. 307–313. 2

[39] T. Fang, X. Zhao, S. K. Shah, and I. A. Kakadiaris, "4d facial expression recognition," in *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, Nov 2011, pp. 1594–1601. 2

[40] P. Yang, Q. Liu, X. Cui, and D. N. Metaxas, "Facial expression recognition using encoded dynamic features," in *2008 IEEE Conference on Computer Vision and Pattern Recognition*, June 2008, pp. 1–8. 2

[41] L. Zhong, Q. Liu, P. Yang, J. Huang, and D. N. Metaxas, "Learning multiscale active facial patches for expression analysis," *IEEE Transactions on Cybernetics*, vol. 45, no. 8, pp. 1499–1510, Aug 2015. 2

[42] K. Zhao, W. S. Chu, F. D. la Torre, J. F. Cohn, and H. Zhang, "Joint patch and multi-label learning for facial action unit and holistic expression recognition," *IEEE Transactions on Image Processing*, vol. 25, no. 8, pp. 3931–3946, Aug 2016. 2

[43] D. Kollias, A. Tagaris, and A. Stafylopatis, "On line emotion detection using retrainable deep neural networks," in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*, Dec 2016, pp. 1–8. 2

[44] M. Rana, N. Gupta, J. L. Dalboni da rocha, S. Lee, and R. sitaram, "A toolbox for real-time subject-independent and subject-dependent classification of brain states from fmri signals," *Frontiers in Neuroscience*, vol. 7, p. 170, 2013. [Online]. Available: http://journal.frontiersin.org/article/10.3389/fnins.2013.00170 2

[45] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *2013 IEEE Conference on Computer Vision and Pattern Recognition*, June 2013, pp. 3444–3451. 2

[46] H. Wang, D. Oneata, J. Verbeek, and C. Schmid, "A robust and efficient video representation for action recognition," *International Journal of Computer Vision*, vol. 119, no. 3, pp. 219–238, Jul. 2015. [Online]. Available: https://hal.inria.fr/hal-01145834 3, 4

[47] F. Perronnin, "Universal and adapted vocabularies for generic visual categorization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 30, no. 7, pp. 1243–1256, July 2008. 4

[48] N. Inoue and K. Shinoda, "A fast and accurate video semantic-indexing system using fast map adaptation and gmm supervectors," *IEEE Transactions on Multimedia*, vol. 14, no. 4, pp. 1196–1205, Aug 2012. 4

[49] N. Dehak, P. J. Kenny, R. Dehak, P. Dumouchel, and P. Ouellet, "Front-end factor analysis for speaker verification," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 19, no. 4, pp. 788–798, May 2011. 4, 5

[50] P. Kenny, G. Boulianne, and P. Dumouchel, "Eigenvoice modeling with sparse training data," *IEEE transactions on speech and audio processing*, vol. 13, no. 3, pp. 345–354, 2005. 5

[51] S. J. Prince and J. H. Elder, "Probabilistic linear discriminant analysis for inferences about identity," in *2007 IEEE 11th International Conference on Computer Vision*. IEEE, 2007, pp. 1–8. 5

[52] D. Chen, X. Cao, L. Wang, F. Wen, and J. Sun, "Bayesian face revisited: A joint formulation," in *European Conference on Computer Vision*. Springer, 2012, pp. 566–579. 5

[53] S. Ioffe, *Probabilistic Linear Discriminant Analysis*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2006, pp. 531–542. 5

[54] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, May 2015, pp. 1–6. 6

[55] A. Yao, J. Shao, N. Ma, and Y. Chen, "Capturing au-aware facial features and their latent relations for emotion recognition in the wild," in *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, ser. ICMI '15. New York, NY, USA: ACM, 2015, pp. 451–458. [Online]. Available: http://doi.acm.org/10.1145/2818346.2830585 6, 8

[56] S. E. Kahou, C. Pal, X. Bouthillier, P. Froumenty, Ç. Gülçehre, R. Memisevic, P. Vincent, A. Courville, Y. Bengio, R. C. Ferrari *et al.*, "Combining modality specific deep neural networks for emotion recognition in video," in *Proceedings of the 15th ACM on International conference on multimodal interaction*. ACM, 2013, pp. 543–550. 6, 8

[57] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "C3D: generic features for video analysis," *CoRR*, vol. abs/1412.0767, 2014. [Online]. Available: http://arxiv.org/abs/1412.0767 8

[58] A. Dapogny, K. Bailly, and S. Dubuisson, "Dynamic facial expression recognition by joint static and multi-time gap transition classification," in *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, vol. 1, May 2015, pp. 1–6. 8

[59] ——, "Pairwise conditional random forests for facial expression recognition," in *2015 IEEE International Conference on Computer Vision (ICCV)*, Dec 2015, pp. 3783–3791. 8

[60] S. Pini, O. Ben-Ahmed, M. Cornia, L. Baraldi, R. Cucchiara, and B. Huet, "Modeling multimodal cues in a deep learning-based framework for emotion recognition in the wild," in *19th ACM International Conference on Multimodal Interaction*, 2017. 8

[61] M. Liu, S. Shan, R. Wang, and X. Chen, "Learning expressionlets on spatio-temporal manifold for dynamic facial expression recognition," in *2014 IEEE Conference on Computer Vision and Pattern Recognition*, June 2014, pp. 1749–1756. 8

[62] J. Chen, T. Takiguchi, and Y. Ariki, "Rotation-reversal invariant hog cascade for facial expression recognition," *Signal, Image and Video Processing*, vol. 11, no. 8, pp. 1485–1492, Nov 2017. [Online]. Available: https://doi.org/10.1007/s11760-017-1111-x 8

Nazil Perveen received the BE degree in computer science and engineering from Guru Ghasidas University (India) in 2009 and the M.Tech degree in computer technology from the National Institute of Technology Raipur (India) in 2012. She is currently working toward the Ph.D degree in computer science and engineering at the Indian Institute of Technology Hyderabad (India). Her research interests include pattern recognition, deep learning, human behavior analysis, emotion recognition and medical image analysis.

Debaditya Roy (S'15) received B.Tech from West Bengal University of Technology (India) in 2011, M.Tech from the National Institute of Technology Rourkela (India) in 2013, and Ph.D from Indian Institute of Technology Hyderabad (India) in 2018. His research interests include action recognition and surveillance analysis.

Dr. C. Krishna Mohan received the BScEd degree from the Regional Institute of Education (India) in 1988, the MCA degree from the S. J. College of Engineering (India) in 1991, the MTech degree in system analysis and computer applications from the National Institute of Technology Surathkal (India) in 2000, and the Ph.D degree in computer science and engineering from the Indian Institute of Technology Madras (India) in 2007. He is currently an associate professor in the Department of Computer Science and Engineering, Indian Institute of Technology Hyderabad (India). His research interests include video content analysis, pattern recognition, and neural networks. He is a member of the IEEE.