Facial Expression Recognition in Videos using Dynamic Kernels

Nazil Perveen, Debaditya Roy, and C. Krishna Mohan

***Abstract*—Recognition of facial expressions across various actors, contexts, and recording conditions in real-world videos involves identifying local facial movements. Hence, it is important to discover the formation of expressions from local representations captured from different parts of the face. So in this paper, we propose a dynamic kernel-based representation for facial expressions that assimilates facial movements captured using local spatio-temporal representations in a large universal Gaussian mixture model (uGMM). These dynamic kernels are used to preserve local similarities while handling global context changes for the same expression by utilizing the statistics of uGMM. We demonstrate the efficacy of dynamic kernel representation using three different dynamic kernels, namely, explicit mapping based, probability-based, and matching-based, on three standard facial expression datasets, namely, MMI, AFEW, and BP4D. Our evaluations show that probability-based kernels are the most discriminative among the dynamic kernels. However, in terms of computational complexity, intermediate matching kernels are more efficient as compared to the other two representations.**

***Index Terms*—Expression recognition, feature extraction, universal attribute model, MAP adaptation, factor analysis, Gaussian mixture model, Fisher kernel, supervector kernel, mean interval kernel, intermediate matching kernel**

I. Introduction

Facial expressions are considered to be one of the most important ways of conveying emotions. In general, facial expressions are mainly recognized through various facial movements. The first approach to track facial movements was facial electromyography (fEMG) which uses electrodes attached to the facial area [1]. However, such an invasive method limits facial mobility and hence, a set of rules were devised to measure the facial movements unobtrusively [2]. These rules are known as the Facial Action Coding System (FACS) which is composed of different facial action units (FAU) where each FAU corresponds to a particular facial movement. FACS has been widely used to categorize different types of facial muscle movements that are considered to be physical expressions of emotions. However, FACS does not capture the different variations that may arise in facial movements and thus other facial features like appearance features, geometric features, etc., have been investigated both spatially and temporally for better understanding of facial expressions [3].

Pantic et al. [4] proposed many approaches for better understanding of facial expressions by capturing local and global features. Multiple features like the histogram of oriented gradients [5], Gabor wavelet transformations [6], scale-invariant feature transform (SIFT) [7], active appearance models [8], facial point tracking [9], etc have been explored for facial expression recognition. A combination of geometric features and appearance features are shown substantial improvement in the recognition of facial expressions [8]. Along with these global and local features, other modalities like audio and text (from spoken words) are also used for recognition of emotions. The methodologies described above focus only on datasets where the facial expressions are shot in a controlled environment with no background movement with frontal facial pose [10]. However, such datasets are not representative of real-world human computer interaction systems like driver-fatigue detection for driver-less cars, facial paralysis analysis, pain detection, etc. For real world scenarios, we need to ensure that the facial expression recognition system is pose-independent, subject-independent, context-insensitive, and robust under different capturing conditions.

In order to address the above mentioned challenges, we propose a single model called universal Gaussian mixture model (uGMM) to capture all the local features, which appear in facial expressions across different poses, scales, and subjects in various illumination conditions with dynamic backgrounds. After training the uGMM, the similarity between any two expression clips needs to be calculated based on the distance between the features in the clips and means of the uGMM mixtures. Kernel methods are one of the most popular measures that transform distances to higher dimension in order to enhance separability across classes [11]. However, most of the kernel methods are applicable for static length patterns which hinders for comparing expression clips where the number of local features vary widely. Hence, we employ dynamic kernels [12] that can handle varying length patterns by either converting them into fixed-length patterns (probability based [13] or explicit mapping based kernels [14]) or choosing the best possible combination of local features (matching based kernels [15]).

The base kernel (i.e. similarity measurement) in the dynamic kernels mentioned above is based on the closeness of the local features across two expression clips. In both probability based kernels and explicit mapping based kernels, the posterior probability of every local feature with respect to the uGMM is considered for kernel computation. In case of matching based kernels, the local features that are most similar to the uGMM means are the only ones considered for base kernel computation. This shows that important local structures which contribute to the significant local facial movements are preserved during kernel computation. These significant facial movements are unique to some expressions, for example, brow

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).

lowerer in angry, and pulling of lip corners and cheek raiser in happy [2]. Hence, dynamic kernels are well suited for representing the similarity of facial expressions.

The proposed approach has the following characteristics: (1) the spatio-temporal features are captured from the various facial region to model the dynamics of facial expressions, (2) a universal Gaussian mixture model is constructed to preserve the local structure, and (3) dynamic kernels for variable length is proposed to effectively preserve the local structure and handle the large variation globally. The proposed work explores different types of dynamic kernels to recognize spontaneous (not posed/natural) facial expressions in an unconstrained (uncontrolled/wild) environment. We demonstrate the efficacy of the proposed approach on publicly available benchmark datasets like constrained dataset, namely, Maja and Michel initiative (MMI) facial expression dataset, spontaneous dataset, namely, Binghamton Pittsburgh 4 dimension (BP4D) spontaneous expression dataset, and unconstrained dataset, namely, acted facial expression in wild (AFEW).

The rest of the paper is organized as follows. Section II presents the related work. Section III describes the proposed approach in detail. Experimental results of the proposed approach on the benchmark datasets are discussed in Section IV. Section V gives the conclusion.

II. Related Work

Various approaches that combine generative models with discriminative embedding techniques have been proposed in the literature for facial expression recognition. In this section, the different feature extraction methods used for constructing generative models, in particular, Gaussian mixture models (GMM) are discussed. Also, the related literature on multiple discriminative embedding techniques is summarized with a focus on dynamic kernels.

A. Facial expression recognition

Stefanos et al. [16] proposed the domain adaptation methodology for each and every facial muscle movement. The feature set used by the authors was the local binary pattern (LBP) applied after face registration using contour landmarks removal and affine transformations. The domain adaptation was then applied for various facial action units independently. In total, 12 facial action units were adapted efficiently, which were view and subject independent. This domain adaptation does not require a large dataset, therefore the approach was claimed to be data efficient by the authors. The domain knowledge was then combined with some expert Gaussian process for the final decision. The experiments were conducted on multi-class and multi-label datasets, namely, MultiPie, DISFA, and FERA2015 facial expressions datasets. The authors had also shown the cross datasets evaluation by testing DISFA against FERA2015 and vice-versa. To handle multiple view variations, i.e. from -45 to +45 degrees and to reduce computational complexity involves in capturing the facial muscle movement information, Tariq et al. [17] proposed a supervised Gaussian mixture model (GMM) based modeling of facial expressions. With bag-of-word (BoW) modeling, image descriptors were calculated using soft vector quantization (SVQ) on the GMM mixtures. In order to minimize the reconstruction loss and maximize the data likelihood, the GMM was learned in a supervised manner and a new set of features were obtained known as supervised SVQ (SSVQ).

For better feature representation, the blend of geometrical and appearance based features known as coordinate based features with neutral face subtraction (CBF-NS) was used [18]. The dictionary of the neutral faces was learned through GMM along with the CBF-NS features. The mean and covariance of each GMM component were considered as the dictionary of neutral faces. The movement of the geometric features was tracked after subtracting the best fit neutral face from the dictionary. The most discriminative feature points were then fused with the SIFT features computed around those points to detect the expression using SVM. The datasets, namely, CK+, MMI, and eNTERFACE were used with both similar and cross-dataset experiments to prove the importance of the work.

Non-verbal gestures are also the crucial part of emotion recognition, as during human-human interaction these non-verbal gestures are playing an important role [19]. Initially, the universal background model was used for capturing variations among different gestures across various individuals. Simultaneously, the segmentation model to capture the elements of gesture sequences using HMM is implemented. Both GMM and HMM models were then used for capturing the dynamics of various gestures robustly in the kernel space. The experiment was performed on the self-constructed dataset, which had a certain gesture sequence for happiness, anger, sad, and neutral. They also computed informative gestures like head motion, hand gestures, etc, which were sufficiently good for classifying 3-class or 4-class based facial expressions. Also, various deep learning techniques like convolution neural networks [20], recurrent neural network [9], etc are explored in past few decades for efficient facial expression recognition in frontal pose faces.

B. Dynamic kernels

Similar approaches that combine generative models like Gaussian mixture model (GMM) and hidden Markov model (HMM) has been used for representing varying length data like speech, music, and actions in video clips. One of the most popular approaches for mapping varying length to fixed length representation is dynamic kernels [21]. Lee et al. [14] proposed a probabilistic sequence kernel (PSK), which used the probabilistic measures of Gaussian densities individually rather than in the combined form as a universal background model (UBM). The main goal was to learn discriminative features rather than generative features. The authors conclude that their method has better discrimination ability than any other sequence kernel. The PSK uses both class-independent GMM and class-specific GMM to generalize the models among various speakers and to learn the example specific information, respectively. The PSK is more efficient than GMM-UBM or any other sequence kernel but has higher computational complexity.

Reducing the computational complexity of the probability sequence kernel (PSK), Chang et al. [22] proposed the use

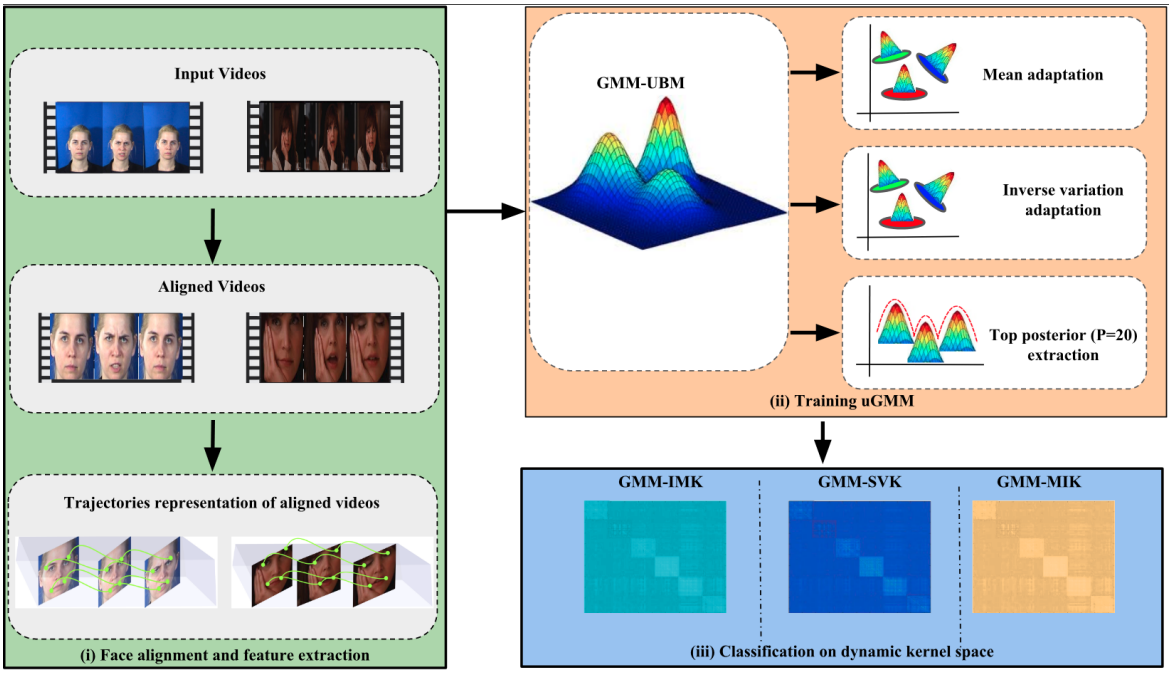

Fig. 1: Block diagram of the proposed spontaneous expression recognition using dynamic kernels.

of Bhattacharyya distance-based measure between two GMM components, which exploits both the first and second order statistics of universal GMM. The authors trained a single generalized background model known as a universal background model (UBM) for modeling features from different speakers. Adapting the means and covariances from the global mean and covariance of the UBM for each speaker resulted in universal mean interval supervectors. The kernel formed using this supervector is called Gaussian mean interval kernel (GUMI) that was used for classification with an SVM. However, GUMI kernel introduces mean and covariance adaptation, which adds the computational load.

In order to further reduce the computational complexity, intermediate matching kernels (IMK) were proposed in [15] that do not require mean or covariance adaptation. Instead, a set of local virtual features based on either GMM mixtures or HMM states was used to select the closest local features from every clip, which were then used for the computation of IMK. The cost of computing IMK was shown to be lower than either GUMI or PSK as the number of virtual features was generally far less than the number of local features extracted from any clip [21]. Further, it was shown that an improvement in computation speed could be achieved by optimizing the selection of virtual features.

To summarize, multiple works were done in the area of facial expression recognition to capture the dynamics of facial movements using various generative and discriminative representations. However, most of the methods like learning-based methods mentioned above, incurs high computational cost with great domain expertise [23].

III. Proposed Approach

Motivated by the literature presented in the previous section, we construct a universal GMM from local spatio-temporal features to represent facial expressions. The statistics of the learned uGMM are then mapped into the dynamic kernel space for efficient facial expression recognition. A number of different dynamic kernels have been studied in this paper and the entire process of kernel formation is shown in Figure 1.

A. Face alignment and feature extraction

The faces of subjects in unconstrained videos are constantly moving across a variety of dynamic backgrounds. Hence, identification and alignment of the subject's face during the entire video clip is crucial for feature extraction. We choose discriminative response map fitting (DRMF) as it has been shown to be both accurate and computationally efficient in the detection of faces [24]. Facial landmark points produced by the DRMF method can be used to eliminate background information in order to create aligned videos of subjects' faces. In these aligned videos, the various facial movements are tracked using a set of densely sampled interest points. For each of these trajectory points, histogram of optical flow (HOF) and motion boundary histogram (MBH) features are calculated [25]. Each expression clip can be represented using the set of feature vectors (either HOF or MBH) $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_L\}$, where L is the number of feature vectors computed from the clip.

B. Training uGMM

After the calculation of the HOF and MBH feature descriptors described above, a separate uGMM is trained for each descriptor. The uGMM is represented by the parameter set $\lambda = \{w_c, \boldsymbol{\mu}_c, \boldsymbol{\sigma}_c\}$ and can be expressed as

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

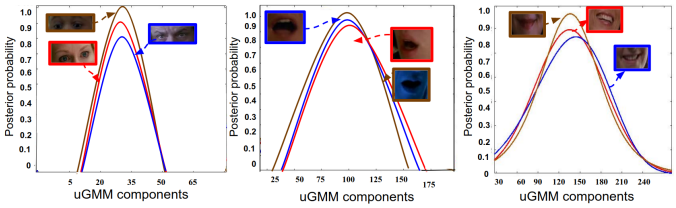

Fig. 2: Each Gaussian component in the uGMM captures a particular facial attribute from different poses and subjects under different illumination conditions. Some examples of facial movements with high posterior probability w.r.t. to a Gaussian component ((a) 30, (b) 100, and (c) 150) are given. A 256 component uGMM trained on AFEW dataset has been considered for this illustration (best viewed in color)

.

where c represents each uGMM component, C represents total uGMM components, and the mixture weights w_c should satisfy the constraint $\sum_{c=1}^{C} w_c = 1$. Also, $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$ are the mean and covariance for component c of the uGMM, respectively. The parameter λ of uGMM is estimated using standard EM method. The EM estimation is an iterative process of finding the maximum likelihood estimate of the model parameters, in this case, uGMM means, covariances, and mixing coefficients. In the E-step, we calculate the membership probabilities of the uGMM mixtures given the features and in the M-step re-estimating and maximizing the parameters using the membership probabilities [26]. We hypothesize that after training the uGMM, each Gaussian component captures a facial motion attribute as shown in Figure 2. Further, the variance of each Gaussian component accounts for variability in poses and subjects under different illumination conditions for each facial motion attribute. These attributes can be specific to a particular expression or may be present in multiple facial expressions. Furthermore, capturing such a large number of attributes in the uGMM allows the comparison of expression clips across a variety of facial movements, which reduces the effect of intra-expression variability.

As the uGMM contains attributes from different facial expressions, representing a particular clip requires enhancing the contribution of the attributes present in the clip. This can be achieved by maximum *aposteriori* (MAP) adaptation. The first step in MAP adaptation is to calculate the probabilistic alignment of every feature vector from a clip with respect to the each mixture of the uGMM that is computed as

$$p(c|\mathbf{x}_l) = \frac{w_c p(\mathbf{x}_l|c)}{\sum_{c=1}^{C} w_c p(\mathbf{x}_l|c)}, \tag{2}$$

where $\mathbf{x}_l$ is a feature vector of the clip and $p(\mathbf{x}_l|c)$ is the likelihood of a feature $\mathbf{x}_l$ arrives from a mixture c. Using the probabilistic alignment, different dynamic kernels can be computed for better representation of variable length patterns to fixed length pattern as described subsequently.

C. Dynamic Kernels

1) Explicit mapping based dynamic kernel: In order to calculate dynamic kernels, the set of local feature vectors are first mapped to fixed length representations and then a kernel function is designed in that space. Using probabilistic alignment computed in Equation 2, the gradient of log-likelihood w.r.t. to the uGMM means is obtained as

$$\boldsymbol{\psi}_c^{(\boldsymbol{\mu})}(\mathbf{X}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{a}_{lc}, \tag{3}$$

where $\mathbf{a}_{lc} = \sum_c^{-1}(\mathbf{x}_l - \boldsymbol{\mu}_c)$. The gradient vector for covariances of each component of the uGMM is given by

$$\boldsymbol{\psi}_c^{(\boldsymbol{\sigma})}(\mathbf{X}) = \frac{1}{2}\left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)\left[-\mathbf{u}_c + \mathbf{v}_{lc}\right]\right), \tag{4}$$

where $\mathbf{u}_c = \Sigma_c^{-1}$ and $\mathbf{v}_{lc} = \left[a_{l1c}\mathbf{a}_{lc}^T, a_{l2c}\mathbf{a}_{lc}^T, \cdots, a_{ldc}\mathbf{a}_{lc}^T\right]$. Finally, the gradient of the log-likelihood w.r.t. the weights is computed as

$$\boldsymbol{\psi}_c^{(w)}(\mathbf{X}) = \sum_{l=1}^{L} p(c|\mathbf{x}_l)\left[\frac{1}{w_c} - \frac{p(c_1|\mathbf{x}_l)}{w_1 p(c|\mathbf{x}_l)}\right]. \tag{5}$$

The gradient of log-likelihood with respect to the parameters of the uGMM defines the directions in which the parameters should be updated to best fit the model. Thus, these gradients show the deviation of the facial motion attribute derived from an expression clip compared to the facial motion attribute captured by the uGMM component. By stacking all the gradients, the Fisher score vector is formed for each component c as

$$\Phi_c(\mathbf{X}) = \left[\boldsymbol{\psi}_c^{(\mu)}(\mathbf{X})^T, \boldsymbol{\psi}_c^{(\sigma)}(\mathbf{X})^T, \boldsymbol{\psi}_c^{(w)}(\mathbf{X})^T\right]^T. \tag{6}$$

The Fisher score with respect to the entire uGMM containing C mixtures is obtained as

$$\Phi_{fs}(\mathbf{X}) = \left[\Phi_1(\mathbf{X})^T\Phi_2(\mathbf{X})^T, \cdots, \Phi_C(\mathbf{X})^T\right]^T. \tag{7}$$

Now, to compare two expression clips, $\mathbf{X}_m$ and $\mathbf{X}_n$ containing L_m and L_n local features, respectively, the Fisher kernel is constructed as

$$\mathbf{K}(\mathbf{X}_m, \mathbf{X}_n) = \Phi_{fs}(\mathbf{X}_m)^T\mathbf{F}^{-1}\Phi_{fs}(\mathbf{X}_n), \tag{8}$$

where $\mathbf{F}$ is the Fisher information matrix expanded as

$$\mathbf{F} = \frac{1}{N}\sum_{f=1}^{N}\Phi_{fs}(\mathbf{X}_f)\Phi_{fs}(\mathbf{X}_f)^T. \tag{9}$$

The Fisher score captures the similarity across two expression clips, whereas the Fisher information matrix captures the variability in the distinct facial movements across two expression clips. The information obtained from these two quantities is combined to form the Fisher kernel given in Equation 8. The construction of Fisher kernel (FK) involves computation of three gradient vectors of dimension $C \times (L_m + L_n)$. Similarly, the computation of Fisher information matrix involves $N \times d_s^2 + N$ computations, where N is the total number of training examples. The final complexity of Fisher kernel computation is $d_s^2 + d_s$, where d_s is the dimension of the Fisher score vector. Hence, the total computation complexity for Fisher kernel is $\mathcal{O}(CL + Nd_s^2 + N + d_s^2 + d_s)$ as shown in Table I.

TABLE I: Computational complexity of different kernels. C is number of uGMM components, L_m and L_n represent the number of local feature vectors for two clips to be compared, d_l is the dimension of the local feature vectors, d_s is the dimension of the score vector, N is the number of training examples.

	Fisher Kernel		Intermediate Matching Kernel		GMM Supervector Kernel		GMM Mean Interval Kernel	
Number of Computations	*Gradient vector computation*	$3 \times C \times (L_m + L_n)$	*Posterior probability computation*	$C \times (L_m + L_n)$	*Mean adaptation*	$C \times (L_m + L_n)$	*Mean adaptation*	$C \times (L_m + L_n)$
	Fisher information matrix computation	$N \times d_s^2 + N$	*Comparisons to select features*	$C \times (L_m + L_n)$	*Supervector computation*	$C \times (d_l^2 + 1)$	*Covariance adaptation*	$C \times (L_m + L_n)$
	Kernel computation	$d_s^2 + d_s$	*Base kernel computation*	C	*Kernel computation*	d_s^2	*Supervector computation*	$C \times (d_l^2 + d_l)$
							Kernel computation	d_s^2
Computational Complexity	$\mathcal{O}(CL + Nd_s^2 + N + d_s^2 + d_s)$		$\mathcal{O}(CL)$		$\mathcal{O}(CL + Cd_l^2 + d_s^2)$		$\mathcal{O}(CL + Cd_l^2 + Cd_l + C^2 d_s^2)$	

2) Probability based dynamic kernel: Instead of gradients, probability based kernels compare the probabilistic distributions of the local feature vectors of two clips. This requires the MAP adapted means and covariances of the uGMM for every clip, that are computed as

$$\boldsymbol{\mu}_c(\mathbf{X}) = \alpha \mathbf{F}_c(\mathbf{X}) + (1-\alpha)\boldsymbol{\mu}_c. \tag{10a}$$

and

$$\boldsymbol{\sigma}_c(\mathbf{X}) = \alpha \mathbf{S}_c(\mathbf{X}) + (1-\alpha)\boldsymbol{\sigma}_c. \tag{10b}$$

Here, $\mathbf{F}_c(\mathbf{X})$ and $\mathbf{S}_c(\mathbf{X})$ denote first and second-order Baum-Welch statistics for a clip $\mathbf{X}$, respectively. These statistics can be calculated as

$$\mathbf{F}_c(\mathbf{X}) = \frac{1}{n_c(\mathbf{X})} \sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{x}_l \tag{11a}$$

and

$$\mathbf{S}_c(\mathbf{X}) = diag\left(\sum_{l=1}^{L} p(c|\mathbf{x}_l)\mathbf{x}_l\mathbf{x}_l^T\right), \tag{11b}$$

respectively. The adaptation of the mean and covariance of every uGMM mixture is based on the posterior probability of that uGMM mixture given the expression clip. A higher posterior probability shows close correlation between the facial motion attribute captured by the Gaussian component to the facial motion attribute in the expression clip. This further implies that adapted means and covariances for that mixture will have larger influence from the first and second-order Baum-Welch statistics $\mathbf{F}_c(\mathbf{X})$ and $\mathbf{S}_c(\mathbf{X})$, respectively than the original uGMM mean and covariance $\boldsymbol{\mu}_c$ and $\boldsymbol{\sigma}_c$, respectively.

Using the adapted means from Equation 10a, the uGMM vector $\boldsymbol{\psi}_c(\mathbf{X})$ for a clip $\mathbf{X}$ is obtained as

$$\boldsymbol{\psi}_c(\mathbf{X}) = \left[\sqrt{w_c}\boldsymbol{\sigma}_c^{-\frac{1}{2}}\boldsymbol{\mu}_c(\mathbf{X})\right]^T. \tag{12}$$

By concatenating the uGMM vectors, a $(Cd \times 1)$-dimensional uGMM supervector (GSV) is obtained for each clip as $\mathbf{s}_{GSV}(\mathbf{X}) = [\boldsymbol{\psi}_1(\mathbf{X})^T, \boldsymbol{\psi}_2(\mathbf{X})^T, \cdots, \boldsymbol{\psi}_C(\mathbf{X})^T]^T$. The supervector kernel between two clips $\mathbf{X}_m$ and $\mathbf{X}_n$ is then given by

$$K_{GSV}(\mathbf{X}_m, \mathbf{X}_n) = \mathbf{s}_{GSV}(\mathbf{X}_m)^T \mathbf{s}_{GSV}(\mathbf{X}_n). \tag{13}$$

Though the supervector takes into account the first order statistics, it does not utilize the second order statistics. So, in order to capture both the second order statistics and deviation of the adapted means from the means of the uGMM, a mean interval vector can be computed for each mixture c of the uGMM as

$$\boldsymbol{\psi}_c(\mathbf{X}) = \left(\frac{\boldsymbol{\sigma}_c(\mathbf{X}) - \boldsymbol{\sigma}_c}{2}\right)^{-\frac{1}{2}} \left(\boldsymbol{\mu}_c(\mathbf{X}) - \boldsymbol{\mu}_c\right). \tag{14}$$

The variability among the adapted parameters and the uGMM components depends on both mean and covariance statistical dissimilarity. Hence, the mean interval vector contains covariance statistical dissimilarity as shown in the first term of Equation 14 and mean statistical dissimilarity which is evaluated in the second term of Equation 14. Combining these mean interval vectors across the uGMM mixtures gives the uGMM mean interval (GMI) supervector $\mathbf{s}_{GMI}(\mathbf{X}) = [\boldsymbol{\psi}_1(\mathbf{X})^T, \boldsymbol{\psi}_2(\mathbf{X})^T, \cdots, \boldsymbol{\psi}_C(\mathbf{X})^T]^T$. Finally, the GMI kernel between two clips $\mathbf{X}_m$ and $\mathbf{X}_n$ is computed as

$$K_{GMI}(\mathbf{X}_m, \mathbf{X}_n) = \mathbf{s}_{GMI}(\mathbf{X}_m)^T \mathbf{s}_{GMI}(\mathbf{X}_n). \tag{15}$$

The construction of uGMM-SVK involves mean adaptation that requires $C \times (L_m + L_n)$ computations and uGMM-MIK involves both mean and covariance adaptation that requires $2 \times C \times (L_m + L_n)$. Similarly, supervector and kernel construction of uGMM-SVK and uGMM-MIK needs $C \times (d_l^2 + 1)$ and d_s^2 computations, respectively, where d_l is the dimension of the local feature vectors. The total computational complexity of uGMM-SVK is $\mathcal{O}(CL + Cd_l^2 + d_s^2)$ and uGMM-MIK is $\mathcal{O}(CL + Cd_l^2 + Cd_l + C^2 d_s^2)$, as shown in Table I.

3) Matching based dynamic kernel: Apart from the explicit mapping and probabilistic based dynamic kernels described above, there are matching based approaches like matching kernel (MK) which compare clips directly based on local similarity across features [13]. Specifically, the matching kernel is constructed by considering the closest local features within a pair of clips as

$$K_{MK}(\mathbf{X}_m, \mathbf{X}_n) = \sum_{l=1}^{L_m} \max_{l'} k(\mathbf{x}_{ml}, \mathbf{x}_{nl'}) + \sum_{l'=1}^{L_n} \max_{l} k(\mathbf{x}_{ml}, \mathbf{x}_{nl'}), \tag{16}$$

where $k(.,.)$ is a base kernel (Gaussian kernel), L_m and L_n represent the number of features in the clips $\mathbf{X}_m$ and $\mathbf{X}_n$, respectively. However, matching kernel is computationally expensive as the number of base kernels to be computed is $\mathcal{O}(L^2)$ where L represents the maximum of L_m and L_n.

To reduce the computational complexity of MK, intermediate matching kernel (IMK) is proposed as an alternative that is constructed by matching sets of feature vectors that are closest to a set of fixed virtual feature vectors. Let $\mathbf{V} = \{\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_C\}$ be the set of virtual feature vectors. Then, the feature vectors that are nearest to the c^{th} virtual feature vector $\mathbf{v}_c$ from clips $\mathbf{X}_m$ and $\mathbf{X}_n$ are determined as

$$\mathbf{x}^*_{mc} = \arg\min_{\mathbf{x}\in\mathbf{X}_m} \mathcal{D}(\mathbf{x}, \mathbf{v}_c) \text{ and } \mathbf{x}^*_{nc} = \arg\min_{\mathbf{x}\in\mathbf{X}_n} \mathcal{D}(\mathbf{x}, \mathbf{v}_c). \tag{17}$$

The distance function $\mathcal{D}(.,.)$ measures the closeness of a feature vector in $\mathbf{X}_m$ or $\mathbf{X}_n$ to a virtual feature vector in $\mathbf{V}$. This distance function allows us to find the best facial movement from each expression clip that matches the facial movement learned for a particular uGMM component. Each uGMM component provides a point of comparison between two expression clips. Hence, having a large number of uGMM components ensures that even small correlations in the facial movement across two expression clips can be measured accurately. This is especially beneficial to resolve high intra-expression variability.

After the selection of the closest feature vectors, a base kernel is computed for each of the C pairs. The IMK is then obtained as the sum of all the C base kernel computations as

$$K_{IMK}(\mathbf{X}_m, \mathbf{X}_n) = \sum_{c=1}^{C} k(\mathbf{x}_{mc}, \mathbf{x}_{nc}). \tag{18}$$

The components of the uGMM, which contains the information in the mean vectors, covariance matrices, and the weights of the uGMM are considered as the virtual feature vectors. As a closeness measure, we use probabilistic alignment, i.e., the posterior probability of a component in the uGMM generating a feature described in Equation 2. So, the local feature vectors close to a particular virtual feature vector (component) c, given by $\mathbf{x}^*_{mc}$ and $\mathbf{x}^*_{nc}$ for clips $\mathbf{X}_m$ and $\mathbf{X}_n$, respectively, are chosen as

$$\mathbf{x}^*_{mc} = \arg\max_{\mathbf{x}\in\mathbf{X}_m} p(c|\mathbf{x}) \text{ and } \mathbf{x}^*_{nc} = \arg\max_{\mathbf{x}\in\mathbf{X}_n} p(c|\mathbf{x}). \tag{19}$$

The construction of IMK involves (i) $C \times (L_m + L_n)$ computations of posterior probabilities for each component, (ii) $C \times (L_m + L_n)$ comparisons to select the closest feature vectors, and (iii) C base kernel computations. This gives a total computational complexity of $\mathcal{O}(CL)$ where L is the maximum of L_m and L_n. When C is smaller than L_m and L_n, the computation is significantly less expensive than other matching kernel as shown in Table I.

IV. Experiments

In this section, the various dynamic kernels are evaluated on different datasets and features. A detailed experimental analysis with multiple features and uGMM components are performed on a variety of expression datasets as shown in Figure 3. In increasing order of complexity, these datasets are: (i) MMI constrained facial expression dataset, (ii) BP4D spontaneous dataset that is recorded in a controlled setup but contains multiple pose and view variations, and (iii) AFEW unconstrained facial expression dataset that is compiled from videos shot in unconstrained environments with various background, pose, and view variations.

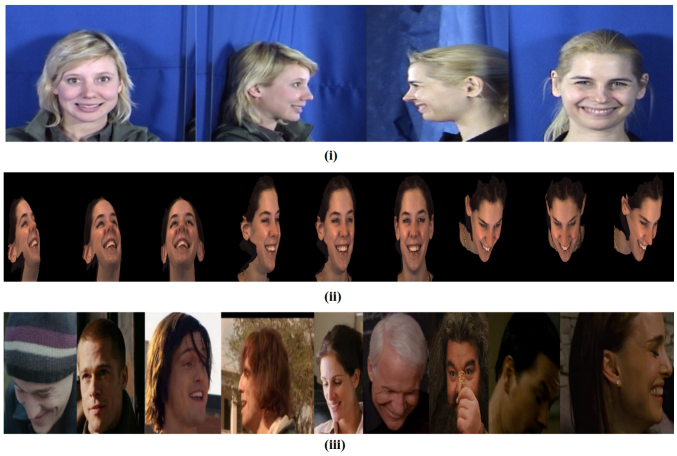

Fig. 3: Facial expression dataset used for the experiment in the proposed approach where (i) belongs to the constraint dataset collected in the laboratory environment, (ii) belongs to the spontaneous dataset collected while interviewing the candidates, (iii) belongs to the unconstrained dataset collected from the movies.

A. Datasets and Experimental Settings

MMI- Maja and Michel Initiative facial expression dataset [10] is a large corpora of audio visual data available for facial expression analysis. The dataset is recorded from onset to the apex to the offset for multiple facial emotions in a constrained environment with acted facial expressions. A total of 75 subjects between the ages of 19 to 62 are recorded in 2900 videos with annotated facial action units (AU) and emotions. However, even with such a large corpora, we were able to experiment with only 213 sequences as these are the only available videos with emotion annotations, in which 180 video sequences are used for training and remaining 33 video sequences are used for testing and a person-independent 10-fold cross-validation is performed as mentioned in [23].

BP4D- The Binghamton Pittsburgh 4D (BP4D) dataset [27] is the first spontaneous facial expression dataset where the facial expressions of the subjects are recorded as a response to videos that invoke a variety of emotions. The dataset contains 2952 videos recorded from 9 different views in a controlled lab environment. A total of 41 subjects participated in the study that includes both male and female of different age groups across different countries. For BP4D, we randomly select videos of 39 subjects for training samples and 2 subjects for testing samples. Though we have trained the uGMM with the 8 different facial expressions of emotions available in the dataset, the reported results include only the 6 emotions,

namely, anger, disgust, fear, happiness, sadness, and surprise. This is done so as to facilitate comparison of our proposed approach with many existing studies [28], [29], [30], [5] where only these 6 facial expressions have been considered.

AFEW- Acted Facial Expression in the Wild (AFEW) [31] is one of the widely used unconstrained facial expression datasets. These datasets are created from Hollywood movies that contain dynamic short videos, which are close to the real-world environment. It consist of total 723 training videos and 383 testing videos. The AFEW dataset is broadly categorized into seven emotion categories, namely, angry, disgust, fear, happiness, neutral, sadness, and surprise.

As mentioned in Section III-A, dense trajectory descriptors - HOF and MBH, which define spatial and temporal characteristics are extracted from each video for the uGMM training. The temporal length of the trajectories used for the computation of HOF and MBH are chosen to be 15 frames as it was found to adequately capture almost all facial movements in their entirety. The HOF descriptor is computed within a space-time volume of size $2 \times 2 \times 3$ and the resulting optical flow is quantized into 9 bins. This results in a final descriptor of 108 dimensions (i.e., $2 \times 2 \times 3 \times 9$). Similarly, MBHx and MBHy (MBH in horizontal and vertical directions, respectively) are computed in the same volume as HOF. Both MBHx and MBHy are quantized into 8 bins leading to a 96-dimensional representation (i.e., $2 \times 2 \times 3 \times 8$) in either direction which are concatenated to form a 192-dimensional descriptor. The computation of the histogram of optical flow (HOF) and motion boundary histogram (MBH) features is based on the trajectories describing facial movement. However, not all trajectories contain information about the facial movements. Some trajectories have sudden and large displacements, which are caused due to the effect of abrupt camera motion. Also, the local constant camera motion like pan, tilt, or zoom is detected by homography estimation and such trajectories are removed to retain only the essential foreground trajectories caused by facial movements. Hence, the HOF and MBH calculated on the foreground trajectories are capable to handle real-world recording conditions in unconstrained environments. The dense trajectory features used in the proposed work compute descriptors within a space-time volume aligned with a trajectory to encode the appearance and motion information, however, the commonly used features for expression recognition like HOG3D, 3DSIFT, and LBP-TOP are usually computed only in a 3D video volume around interest points while ignoring the fundamental dynamic structures in the video [32]. Hence, we use dense trajectory descriptors, namely, HOF and MBH to train 8 uGMMs, i.e. two for each four different 64, 128, 256, and 512 Gaussian components on 3 datasets, namely, MMI, BP4D, and AFEW, respectively. It is observed that the increase in the number of mixtures further, from 256 to 512 does not yield any improvement in classification accuracy of the kernel-based representation due to the increased demand for local features which cannot be met by these datasets.

For adaptation of the uGMM parameters with respect to a video, we consider the top 20 mixtures from the uGMM based on highest posterior probability of the mixture with respect to the local features in the video. Then, the corresponding means and covariance of these mixtures are adapted according to Equations 10a and 10b, respectively. The reason for choosing the top 20 mixtures is that the posterior probabilities of mixtures beyond the first 20 mixtures is found to be mostly zero, which does not influence kernel computation.

B. Comparison of dynamic kernels

The various types of kernels are compared in terms of classification performance in Tables II, III, and IV on the MMI, BP4D, and AFEW datasets, respectively. The classifier used for evaluation is kernel based formulation of support vector machines (SVM) using LibSVM [33]. It can be observed that increasing the number of mixtures in uGMM improves the classification performance of all kernel methods, however, as mentioned earlier the classification performance does not improve beyond 256 Gaussian components. As more uGMM components represented subtle motion dynamics of the face, the unique characteristics can be better captured leading to an improvement in discrimination ability. Further, probability based kernels such as SVK and MIK perform better than matching and mapping based kernels, i.e. IMK and FK. This shows that incorporating the first-order and second-order statistics of the uGMM in probabilistic kernels provides global contextual information that helps in dealing with pose, actor, and illumination variations. Such complementary information augments the local feature based kernel representation that reduces misclassification when compared to matching or mapping based kernels. However, the calculation of first and second-order statistics for uGMM-MIK leads to a computational load far greater than IMK. This trade-off between accuracy and speed can be considered when choosing the appropriate kernel for specific use-cases. Further, MBH is the stable derivative of the the HOF and thus perform better across the facial expression datasets and kernels. As the expression videos contain little camera motion or sudden movements, stabilization does not particularly contribute to classification performance.

TABLE II: Classification performance (in %) of dynamic kernels IMK, FK, SVK, and MIK on MMI

Number of uGMM components	IMK		FK		SVK		MIK	
	HOF	MBH	HOF	MBH	HOF	MBH	HOF	MBH
64	29.5	36.4	38.6	56.8	34.6	42.5	52.3	55.2
128	36.4	40.9	37.3	50.7	45.5	45.9	69.1	69.2
256	38.6	56.8	36.4	45.5	53.3	57.5	71.3	**73.2**
512	37.2	45.4	34.2	42.7	46.8	52.9	70.2	70.5

TABLE III: Classification performance (in %) of dynamic kernels IMK, FK, SVK, and MIK on BP4D

Number of uGMM components	IMK		FK		SVK		MIK	
	HOF	MBH	HOF	MBH	HOF	MBH	HOF	MBH
64	28	25.1	25.8	24.2	47.1	51.5	70.1	52.7
128	34.3	28.6	21.3	21.5	47.6	53.3	71.19	62.4
256	58.9	58.5	18.8	18.9	50.7	53.5	**74.5**	73.3
512	45.7	49.3	18.8	18.9	47.8	52.4	71.6	70.8

TABLE IV: Classification performance (in %) of dynamic kernels IMK, FK, SVK, and MIK on AFEW.

Number of uGMM components	IMK		FK		SVK		MIK	
	HOF	MBH	HOF	MBH	HOF	MBH	HOF	MBH
64	37.8	44.7	25.8	25.8	45.1	39.1	46.5	49.8
128	38.7	47.2	24	18.8	45.8	41.7	48.3	52.3
256	49.5	53.1	21.4	21.4	47.5	44.8	50.3	**56.9**
512	40.2	48.7	20.13	20.3	45.8	42.7	47.5	54.5

C. Expression-wise analysis

Figure 4 represents the confusion matrix of different facial expressions on BP4D dataset. The best recognition accuracy achieve is 74.5% with 256 components and using uGMM-MIK. It can be observed that the classification performance for each expression is similar to overall classification performance. Also, true positive and false positives follow the same trends on the dataset as in classification performance. This shows the generalization capability of the learned model and the similarities in the capturing of the local features through second order statistics i.e. uGMM-MIK based representations. It can be observed that happy facial expressions are misclassified into angry facial expressions and vice-versa because of the common facial movements like stretching of lips and widening of eyes [1]. On the other hand, there is a clear discrimination between happy and surprise expressions. This may be due to the fact that opening of eyes [1] and mouth are rarely found in happy expression but frequently occur in surprise expression.

The kernel matrix for uGMM-MIK on BP4D dataset is visualized in Figure 5. It can be clearly observed that using MIK as a distance measure provides clear separability across different facial expressions.

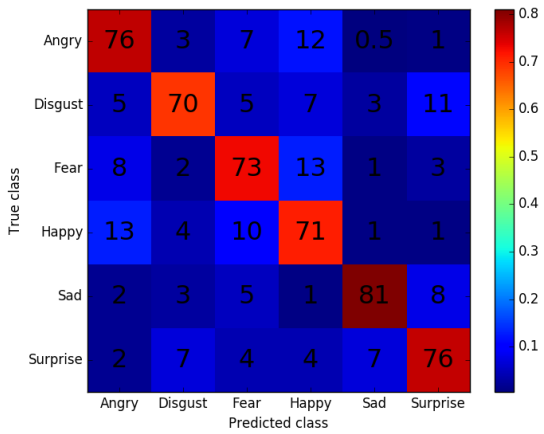

Fig. 4: Confusion matrix (based on classification performance %) of the best model on BP4D dataset, feature extractor used - HOF, Gaussian components - 256, and kernel used - MIK (best viewed in color).

[1]"Widening of eye" represents the positive expression of attention that brings joy and happiness whereas "opening of eye" represents anxiety and strain in the human eye [34].

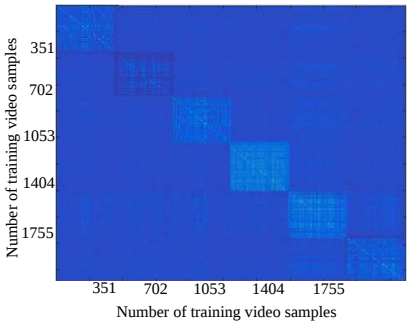

Fig. 5: Kernel matrix representation of the best model on BP4D dataset, feature extractor used - HOF, Gaussian components - 256, and kernel used - MIK. The number on the axis represent clips of different facial expressions (best viewed in color). The diagonal blocks in lighter blue represent higher values as compared to the off diagonal elements in darker blue.

D. Comparison with state-of-the-art

In Tables V, VI, and VII, the proposed method of using dynamic kernels for expression recognition is compared with state-of-the-art approaches on MMI, BP4D, and AFEW datasets, respectively. Existing techniques use low-level features like HOG [5], SIFT [18], and geometric feature [35] that cannot handle the variations encountered in expression videos in unconstrained environments [36]. As these features are extracted at every frame, they consider only spatial information which is not sufficient to analyze facial expressions in videos. By adding temporal information, features like 3D-CNN [28], 3D-CNN on deformable facial parts [37], CNN with conditional random fields (CRF) [30], CNN with bidirectional long short term memory (BLSTM) [29], etc. can adequately represent facial expressions. However, CNNs and LSTMs need large annotated expression datasets in order to generalize well across facial expressions. Hence, the state-of-the-art methods employ a combination of CNN and HOG features with traditional sequential models like HMM.

TABLE V: Performance (%) comparison of dynamic kernels and state-of-the-art methods on MMI dataset

Methods	Accuracy (%)
3DCNN* [28]	53.5
SIFT + SVM [18]	62.5
LDA + NN [38]	67.4
3DCNN + DAP [37]	62.2
Neutral face + sparsity [39]	70.1
CNN + Joint fine-tuning [20]	70.24
Proposed uGMM-FK	56.8
Proposed uGMM-IMK	56.8
Proposed uGMM-SVK	57.5
Proposed uGMM-MIK	**73.2**

* our evaluation using C3D features

In order to further improve the recognition performance, other modalities like audio features are also fused with video

features for integrating complementary information [40]. It can be clearly observed that MIK performs better than the state-of-the-art with the help of the temporal information embedded in the MBH features. The first and second-order statistics of the uGMM used for capturing global context of motion dynamics that are more useful for classification of facial expressions than similar models like deformable facial parts [37] that also capture the structure of local motion dynamics. Also, from Table VI, it can be observed that on the BP4D dataset, the use of higher order statistics from a non-sequential model like uGMM outperforms methods like HMM [5], LSTM [29], and CRF [30] with the classification performance of 74.5% that model facial expressions as sequences. As sequential methods generally require more training data, kernel methods like MIK can be considered as a suitable alternative in absence of large training data.

TABLE VI: Performance (%) comparison of dynamic kernels and state-of-the-art methods on BP4D dataset

Methods	Accuracy (%)
3DCNN* [28]	46.5
CNN + Binary LSTM [29]	54.8
CNN + CRF [30]	66.7
NEBULA + SVM [41]	69.9
HOG + HMM [5]	72.2
Proposed uGMM-FK	25.8
Proposed uGMM-IMK	58.9
Proposed uGMM-SVK	53.5
Proposed uGMM-MIK	**74.5**

* our evaluation using C3D features

TABLE VII: Performance (%) comparison of dynamic kernels and state-of-the-art methods on on AFEW dataset

Methods	Accuracy (%)
3DCNN* [28]	31.3
Expressionlet [42]	31.7
HOG-TOP + geometric warping [35]	45.2
CNN + kernel ELM + PLS [40]	54.5
Proposed uGMM-FK	25.8
Proposed uGMM-IMK	49.5
Proposed uGMM-SVK	47.5
Proposed uGMM-MIK	**56.9**

* our evaluation using C3D features

V. Conclusion

In this paper, we introduce a novel approach for facial expression recognition in videos by using dynamic kernels. Dynamic kernels provide a generic mechanism for incorporating generative models into discriminative classifiers. A universal GMM (uGMM) model with simple kernel computations is used to capture the local dynamics while preserving variations in global context. We have shown that by subsuming first and second order statistics of uGMM, a kernel based representation can be derived for recognizing the facial expressions efficiently. We evaluate the proposed approach on three challenging benchmark datasets to show the generic mechanism of the proposed approach for video-based facial expression recognition. We have also shown that the probability based mean interval kernel (MIK) outperforms other state-of-the-art approaches. Also, IMK are shown to be computationally efficient but is not as discriminative as MIK. This makes dynamic kernels a natural choice for any expression recognition application that focuses either on accuracy or computation time.

References

[1] U. Dimberg, "For distinguished early career contribution to psychophysiology: Award address, 1988," *Psychophysiology*, vol. 27, no. 5, pp. 481–494, 1990. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1469-8986.1990.tb01962.x 1

[2] P. Ekman, J. Hager, and W. Friesen, *The Symmetry of Emotional and Deliberate Facial Actions*. Society for psychophysiological research, Incorporated. [Online]. Available: https://books.google.co.in/books?id=yAyInQAACAAJ 1, 2

[3] I. A. Essa and A. P. Pentland, "Coding, analysis, interpretation, and recognition of facial expressions," *IEEE transactions on pattern analysis and machine intelligence*, vol. 19, no. 7, pp. 757–763, 1997. 1

[4] 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, Jan 2009. 1

[5] 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. 1, 7, 8, 9

[6] 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. 1

[7] 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. 1

[8] S. L. Happy and A. Routray, "Automatic facial expression recognition using features of salient facial patches," *IEEE Transactions on Affective Computing*, vol. 6, no. 1, pp. 1–12, Jan 2015. 1

[9] K. Zhang, Y. Huang, Y. Du, and L. Wang, "Facial expression recognition based on deep evolutional spatial-temporal networks," *IEEE Transactions on Image Processing*, vol. 26, no. 9, pp. 4193–4203, Sep. 2017. 1, 2

[10] M. Pantic, M. Valstar, R. Rademaker, and L. Maat, "Web-based database for facial expression analysis," in *2005 IEEE International Conference on Multimedia and Expo*, July 2005, pp. 5 pp.–. 1, 6

[11] C. J. Burges, "A tutorial on support vector machines for pattern recognition," *Data Mining and Knowledge Discovery*, vol. 2, no. 2, pp. 121–167, Jun 1998. [Online]. Available: https://doi.org/10.1023/A:1009715923555 1

[12] V. Wan and S. Renals, "Svmsvm: support vector machine speaker verification methodology," in *Acoustics, Speech, and Signal Processing, 2003. Proceedings. (ICASSP '03). 2003 IEEE International Conference on*, vol. 2, April 2003, pp. II–221–4 vol.2. 1

[13] C. Wallraven, B. Caputo, and A. Graf, "Recognition with local features: the kernel recipe," in *Proceedings Ninth IEEE International Conference on Computer Vision*, Oct 2003, pp. 257–264 vol.1. 1, 5

[14] K.-A. Lee, C. You, H. Li, and T. Kinnunen, "A gmm-based probabilistic sequence kernel for speaker verification," 2007. 1, 2

[15] S. Boughorbel, J.-P. Tarel, and N. Boujemaa, "The intermediate matching kernel for image local features," in *Proceedings of International Joint Conference on Neural Networks (IJCNN'05)*, Montréal, Canada, 2005, pp. 889 – 894, http://perso.lcpc.fr/tarel.jean-philippe/publis/ijcnn05.html. 1, 3

[16] S. Eleftheriadis, O. Rudovic, M. P. Deisenroth, and M. Pantic, "Gaussian process domain experts for model adaptation in facial behavior analysis," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, June 2016, pp. 1469–1477. 2

[17] U. Tariq, J. Yang, and T. S. Huang, "Maximum margin gmm learning for facial expression recognition," in *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, April 2013, pp. 1–6. 2

[18] S. Ulukaya and C. E. Erdem, "Gaussian mixture model based estimation of the neutral face shape for emotion recognition," *Digital Signal Processing*, vol. 32, pp. 11 – 23, 2014. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1051200414001651 2, 8

[19] Z. Yang and S. S. Narayanan, "Modeling dynamics of expressive body gestures in dyadic interactions," *IEEE Transactions on Affective Computing*, vol. 8, no. 3, pp. 369–381, July 2017. 2

[20] H. Jung, S. Lee, J. Yim, S. Park, and J. Kim, "Joint fine-tuning in deep neural networks for facial expression recognition," in *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*, ser. ICCV '15. Washington, DC, USA: IEEE Computer Society, 2015, pp. 2983–2991. [Online]. Available: http://dx.doi.org/10.1109/ICCV.2015.341 2, 8

[21] A. D. Dileep and C. C. Sekhar, "Gmm-based intermediate matching kernel for classification of varying length patterns of long duration speech using support vector machines," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 25, no. 8, pp. 1421–1432, Aug 2014. 2, 3

[22] C. H. You, K. A. Lee, and H. Li, "Gmm-svm kernel with a bhattacharyya-based distance for speaker recognition," *IEEE Transactions on Audio, Speech, and Language Processing*, vol. 18, no. 6, pp. 1300–1312, Aug 2010. 2

[23] 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, 2016. 3, 6

[24] 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. 3

[25] 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

[26] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification (2Nd Edition)*. New York, NY, USA: Wiley-Interscience, 2000. 4

[27] 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, best of Automatic Face and Gesture Recognition 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885614001012 6

[28] 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. 7, 8, 9

[29] S. Jaiswal and M. Valstar, "Deep learning the dynamic appearance and shape of facial action units," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, March 2016, pp. 1–8. 7, 8, 9

[30] B. Hassani and M. H. Mahoor, "Spatio-temporal facial expression recognition using convolutional neural networks and conditional random fields," *CoRR*, vol. abs/1703.06995, 2017. 7, 8, 9

[31] 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. 7

[32] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu, "Dense trajectories and motion boundary descriptors for action recognition," *International Journal of Computer Vision*, vol. 103, no. 1, pp. 60–79, May 2013. [Online]. Available: https://doi.org/10.1007/s11263-012-0594-8 7

[33] C.-C. Chang and C.-J. Lin, "Libsvm: a library for support vector machines," *ACM transactions on intelligent systems and technology (TIST)*, vol. 2, no. 3, p. 27, 2011. 7

[34] P. Ekman and M. OSullivan, "Who can catch a liar?" *American Psychologist*, vol. 46, no. 9, pp. 913–920, 1991. [Online]. Available: https://doi.org/10.1037/0003-066x.46.9.913 8

[35] J. Chen, Z. Chen, Z. Chi, and H. Fu, "Facial expression recognition in video with multiple feature fusion," *IEEE Transactions on Affective Computing*, vol. 9, no. 1, pp. 38–50, Jan 2018. 8, 9

[36] 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. 8

[37] M. Liu, S. Li, S. Shan, R. Wang, and X. Chen, "Deeply learning deformable facial action parts model for dynamic expression analysis," in *Computer Vision – ACCV 2014*, D. Cremers, I. Reid, H. Saito, and M.-H. Yang, Eds. Cham: Springer International Publishing, 2015, pp. 143–157. 8, 9

[38] H. Hussein, M. Naqvi, and J. Chambers, "Study of image-based expression recognition techniques on three recent spontaneous databases," in *2017 22nd International Conference on Digital Signal Processing (DSP)*, Aug 2017, pp. 1–5. 8

[39] S. H. Lee, W. J. Baddar, and Y. M. Ro, "Collaborative expression representation using peak expression and intra class variation face images for practical subject-independent emotion recognition in videos," *Pattern Recognition*, vol. 54, pp. 52 – 67, 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0031320316000108 8

[40] H. Kaya, F. Grpnar, and A. A. Salah, "Video-based emotion recognition in the wild using deep transfer learning and score fusion," *Image and Vision Computing*, vol. 65, pp. 66 – 75, 2017, multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0262885617300367 9

[41] 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. 9

[42] 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. 9