Quantitative Analysis of Facial Paralysis using GMM and Dynamic Kernels

Nazil Perveen[1][a], Chalavadi Krishna Mohan[1][b] and Yen Wei Chen[2][c]
[1]*Department of Computer Science and Engineering, IIT Hyderabad, Hyderabad, India*
[2]*College of Information Science and Engineering, Ritsumeikan University, Kusatsu, Shiga, Japan*

Keywords: Facial Paralysis, Spatial and Temporal Features, Gaussian Mixture Model, Dynamic Kernels, Expression Modeling, Yanagihara Grading Scales.

Abstract: In this paper, the quantitative assessment for facial paralysis is proposed to detect and measure the different degrees of facial paralysis. Generally, difficulty in facial muscle movements determines the degree with which patients are affected by facial paralysis. In the proposed work, the movements of facial muscles are captured using spatio-temporal features and facial dynamics are learned using large Gaussian mixture model (GMM). Also, to handle multiple disparities occurred during facial muscle movements, dynamic kernels are used, which effectively preserve the local structure information while handling the variation across the different degree of facial paralysis. Dynamic kernels are known for handling variable-length data patterns efficiently by mapping it onto a fixed length pattern or by the selection of a set of discriminative virtual features using multiple GMM statistics. These kernel representations are then classified using a support vector machine (SVM) for the final assessment. To show the efficacy of the proposed approach, we collected the video database of 39 facially paralyzed patients of different ages group, gender, and from multiple angles (views) for robust assessment of the different degrees of facial paralysis. We employ and compare the trade-off between accuracy and computational loads for three different categories of the dynamic kernels, namely, explicit mapping based, probability-based, and matching based dynamic kernel. We have shown that the matching based kernel, which is very low in computational loads achieves better classification performance of 81.5% than the existing methods. Also, with the higher-order statistics, the probability kernel involves more communication overhead but gives significantly high classification performance of 92.46% than state-of-the-art methods.

1 INTRODUCTION

Facial paralysis is the facial nerve paralysis, which occurs due to temporary or permanent damage to the facial nerve. There are multiple reasons like surgical, neurological, viral infections, injuries, etc., which causes damage to the facial nerve. Due to the damage in the facial nerve, there is loss in the movement of the facial muscles, which restrain the patients to pose normal facial actions like smiling, closing of eyes, closing of the mouth, etc. Facial paralysis affect the patient face either on half or both sides.

To detect the level and intensity of the effect caused by the facial paralysis to the patients face, multiple diagnoses are required by the clinicians. Most of them involve subjective assessments like assigning of grading score to the patient face based on certain facial expressions. The Yanagihara grading scale by Hato et al. (2014) and House-Brackmann (HB) grading scales by House and Brackmann (1985) are the two mostly used subjective grading scores for evaluating the facial paralysis and its effects. Due to the easier interpretation of the grading levels and the formation of the facial simple expression Satoh et al. (2000), Yanagihara is the widely used techniques in detecting different levels of facial paralysis. There are 10 expressions in Yanagihara grading scale like rest videos (EP0), raising of eyebrows (EP1), closure of eye gently (EP2), closure of eye tightly (EP3), closure of paralyzed eye (EP4), wrinkle nose (EP5), puff out cheeks (EP6), toothy movement (EP7), whistling movement (EP8), and under lip turn down (EP9). Also, there are two different levels of the grading scales using Yanagihara grading rules,

[a] https://orcid.org/0000-0001-8522-7068
[b] https://orcid.org/0000-0002-7316-0836
[c] https://orcid.org/0000-0002-5952-0188

Perveen, N., Mohan, C. and Chen, Y.
Quantitative Analysis of Facial Paralysis using GMM and Dynamic Kernels.
DOI: 10.5220/0009104801730184
In *Proceedings of the 15th International Joint Conference on Computer Vision, Imaging and Computer Graphics Theory and Applications (VISIGRAPP 2020) - Volume 5: VISAPP*, pages 173-184
ISBN: 978-989-758-402-2; ISSN: 2184-4321

i.e, 5-scores, and 3-scores grading scales. In 5-scores grading scales, the listed expression posed by the patient is graded from score-0 to score-4, where score-0 denotes high-level of facial paralysis, score-1 denotes almost facial paralysis, score-2 represents moderate, score-3 represents slight facial paralysis and score-4 denotes no facial paralysis. Similarly, in 3-scores grading scales, the listed expression posed by the patient is graded from score-0 to score-2, where score-0 denotes high-level of facial paralysis and score-2 denotes low-level of facial paralysis (or no paralysis), respectively.

Although subjective assessments are widely used techniques but it highly depends on the expert's opinion of assigning grades while examining the patients during facial expressions formation. This motivates our research to develop a generalized model for the quantitative assessments of the facial paralysis using different dynamic kernels. Kernels effectively preserved the local structure and also able to handle large variation globally. Thus, in the proposed approach once the local attributes are captured implicitly by the components of universal GMM, the kernels are learned for the better representation of the video both, locally and globally. Also, the video data mostly contains a variable length sequence of the local feature vector, therefore to handle the variability in the sequence of local features extracted from the videos, dynamic kernels are used Dileep and Sekhar (2014).

The paper is organized as follows. In Section 2, we discuss the previous work done for the quantitative assessment of facial paralysis. Section 3 describes the proposed quantitative assessment method in detail. Experimental results are discussed in Section 4 to show the efficacy of the proposed approach. Section 5 concludes the work with future directions.

2 RELATED WORK

NGO et al. (2016) proposed the quantitative assessment of the facial paralysis using 2D features. These 2D features were novel and robust spatio-temporal features, which were computed frame-wise. Initially, face was detected in the given frame using the AdaBoost algorithm and then landmarks points were detected. The facial landmark points were detected by computing region of interest (ROI) using the perpendicularity of inter-pupil distance with vertical face mid line. Once the ROI area was selected the landmark points were placed and tracked throughout the frames. The spatio-temporal features were extracted using the tracked landmark points, which were then classified using support vector machine (SVM) for finding a different level of facial paralysis. The average accuracy achieved by this method is approximately 70% for only three categories of expressions for 5-scoring levels.

He et al. (2009), proposes the novel block processing techniques to capture the appearance information at different resolutions. They use local binary pattern (LBP) to extract appearance features from the apex frame (i.e the frame in which facial expression is highly active) at multiple block levels and at different resolutions, which is known as multi-resolution LBP (MLBP). These blocks were centered over the facial regions like eyebrows, eyes, nose, and mouth. They also, extracted motion information by tracking the facial muscle movement in the horizontal (x-axis) and vertical (y-axis) direction. Once the feature from different regions is extracted they compare the symmetry between normal facial regions with the paralyzed facial region using resistor-average distance (RAD). Finally, they use a support vector machine for the final assessment and score prediction based on RAD. They evaluate their model with the House-Brackmann (HB) grading scale on the self-collected and annotated database. They use four expressions with a 5-score grading scale to achieve average classification rate of 86.6%.

Liu et al. (2015) propose the thermal imaging model for learning the facial paralysis effect. The proposed approach demonstrate the change in facial nerve functions, when the facial temperature changes. The medical infrared thermal imager made of liquid nitrogen was used for facial temperature distribution acquisition. For collecting the infrared thermal image dataset, patients should not drink and must sit for 20 minutes prior to adapt the room temperature before experiments start. Using the features like temperature distribution, area ratio, and temperature difference over the region of interest of normal and paralyzed facial area, they classify the different level of facial paralysis. For classification, K-nearest neighbor classifier (K-NN), support vector machines (SVM), and radial basis function neural network (RBFNN) was used. They evaluate their model for four expressions with an average accuracy of 94% with RBFNN classifier.

Banks et al. (2015) developed the offline application named *eFace* for detection of the unilateral facial paralysis. The video of the patient with posing list of facial expressions was recorded and fed into the eFace for comparing the normal and affected side. Different score to capture disfigurement severity was calculated like static scores, dynamic scores, and synkinesis score. Based on the computed scores the grading from 1 to 100 is provided where 1 denotes high dis-

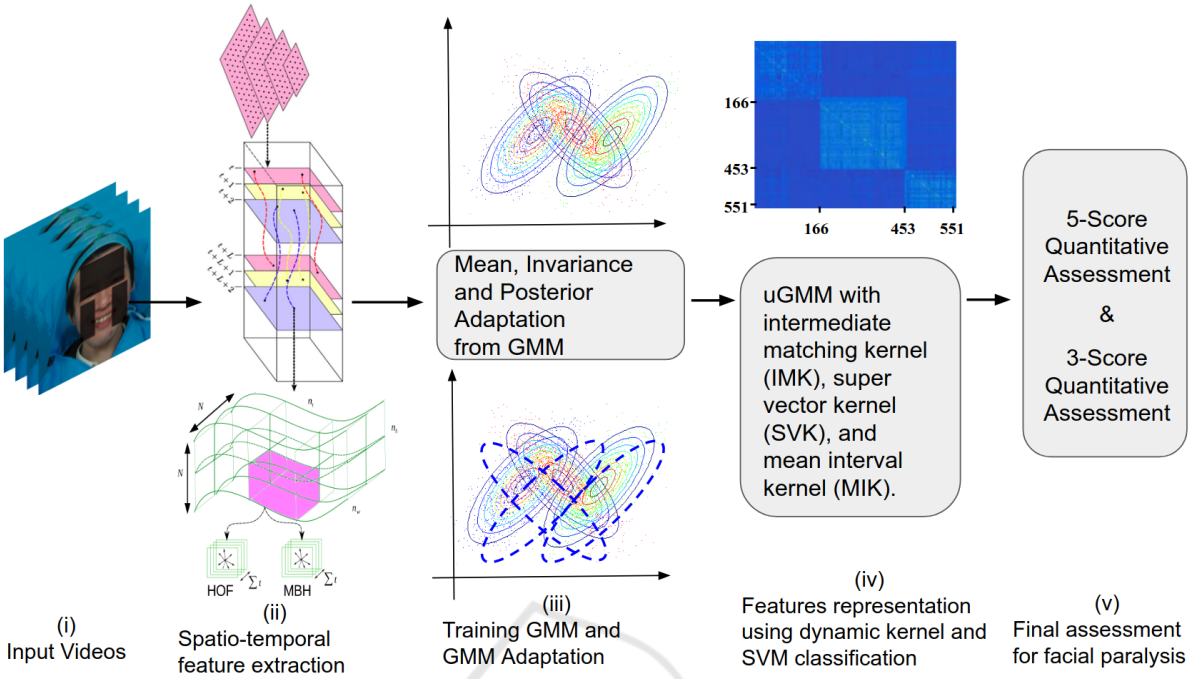


Figure 1: Block diagram of the proposed approach, which comprises of the following steps.

figurement severity and 100 denotes least or no disfigurement severity. The evaluation was done on a self-collected database of 25 subjects under expert supervisions.

Ngo et al. (2016) proposed the objective evaluation of the facial paralysis using 3D features. The facial landmark points were evaluated from the first frame, which was tracked throughout the frames for the given video. These landmarks points were then used to calculate asymmetrically and movement features in the 3-D space for capturing and comparing the facial muscle movement between normal and paralyzed side of the face. This objective evaluation of facial paralysis achieves an average recognition rate of 66.475% for the four prominent expressions, i.e for EP1, EP3, EP5, and EP7, respectively, for 5-scoring levels.

Recently, Guo et al. (2017) proposed the deep neural network model for classifying the severity of the facial paralysis. They use Google LeNet model for the self-collected private database from 104 subjects. They achieved the performance of 91.25% for four expressions with a 5-score House-Brackmann (HB) grading scale.

Thus, the related works mentioned above is highly subjective and most of the approaches are based on asymmetric features. Faces considered in the above approaches are mostly frontal with patient posing very few active expressions like opening and closing of eyes, mouth, etc. This motivates us to develop the generalized model for predicting and classifying different levels of facial paralysis by considering all listed expressions in the literature. The following are the major contribution of the paper:

1. The proposed approach train a large GMM with seven views and multiple subjects to learn view and subject invariant attributes from the videos for the better assessments.
2. The proposed approach introduces dynamic kernels, which handles the variation across various facial muscle movements and effectively preserve the local dynamic structure to distinguish the different degree level of facial paralysis.
3. The proposed approach, model all the 10 expressions mentioned in the Yanagihara grading system with all available scores i.e. 5-scores and 3-score grading scales for effective assessment.

Thus, the propose approach address the limitations of the existing work for better quantitative assessment of facial paralysis. The next section describes the proposed approach in details.

3 PROPOSED APPROACH

The block diagram of the proposed approach is shown in Figure 1. Initially, face in the collected input videos is aligned using facial landmark points to remove unwanted background information. The aligned faces

are then tracked for spatial and temporal feature extraction. These features are then used for training large Gaussian mixture model (GMM), which is then used to compute the statistics of GMM for designing of the dynamic kernel for quantitative assessment of the facial paralysis. The details of the proposed methodology are given as follows.

3.1 Data Pre-processing and Feature Extraction

From the aligned face video, two descriptors, namely, the histogram of optical flow (HOF) and motion boundary histogram (MBH) features are evaluated using dense trajectories Wang et al. (2015) as shown in Figure 1-part (ii). Initially, the dense trajectory feature points are computed at 8 different spatial scales. In each scale, the feature points are densely sampled on a grid spaced by $W = 5$ pixels. Further, each feature points are tracked till the next frame by using median filtering in the dense optical flow field. The trajectories computed are tended to drift from their initial location if tracked for the longer period, thus, to avoid the drifting issue the frame length for tracking is fixed to $t = 15$ frames.

Further, the local descriptors are computed around the interest points in 3D video volume, as it is always the effective way of capturing the motion information. The size of the video volume considered is $P \times P$ pixels, where $P = 32$. To ensure the dynamic structure of the video the volume is further subdivided into spatio-temporal grid of $g_h \times g_w \times g_t$, where $g_h = 2$, $g_w = 2$, and $g_t = 3$ are height, width, and temporal segment lengths. Once, the HOF and MBH descriptors are computed from each spatio-temporal grid, it is quantized into 9 and 8 bins, respectively, and normalized using RootSIFT method as mentioned in Wang et al. (2015).

The size of the HOF descriptors obtained is of 108 dimensions (i.e., $2 \times 2 \times 3 \times 9$). Also, the size of MBH descriptors obtained by computing the descriptors from horizontal and vertical components of the optical flow, i.e MBH in x and y direction are of 192 dimensions (i.e. 96 dimensions for each direction).

The reason for using the above mention dense trajectory, i.e., HOF and MBH descriptors are, all the trajectories in the given video does not contain useful information like trajectories cause due to large, sudden, and constant camera motions. Therefore these trajectories are required to remove so to retain only the essential foreground trajectories caused by the facial movements. The removal of such trajectories is efficiently done by the improved dense trajectories, which are far efficient than commonly used features like HOG3D, 3DSIFT, and LBP-TOP, etc, that are usually computed in a 3D video volume around interest points, which usually ignores the fundamental dynamic structures in the video Wang et al. (2015).

3.2 Training of a Gaussian Mixture Model (GMM)

The features obtained from different views and subjects from the videos are extracted to train the large Gaussian mixture model (GMM). The GMM is trained for multiple components $q = 1, 2, \cdots, Q$ in order to capture different facial movement attribute in various Q components. Given a video $\mathbf{V}$, the set of local features are represented as $\mathbf{v}_1, \mathbf{v}_2, \cdots, \mathbf{v}_N$, where N is the total number of local features for the given V. The likelihood of the particular feature $\mathbf{v}_n$ generated from the GMM model is given by

$$p(\mathbf{v}_n) = \sum_{q=1}^{Q} w_q \mathcal{N}(\mathbf{v}_n | \mu_q, \sigma_q), \tag{1}$$

where μ_q, σ_q represents mean and covariance for each GMM component q, respectively. Further, w_q represents GMM mixture weights, which should satisfy the constraint $\sum_{q=1}^{Q} w_q = 1$. Once the GMM is trained, the probabilistic alignment of each feature vector $\mathbf{v}_n$ with respect to the qth component of the GMM model is evaluated using as follows

$$p(q|\mathbf{v}_n) = \frac{w_q p(\mathbf{v}_n|q)}{\sum_{q=1}^{Q} w_q p(\mathbf{v}_n|q)}, \tag{2}$$

where $p(\mathbf{v}_n|q)$ is the likelihood of a feature $\mathbf{v}_n$ generated from a component q. Using the different parameters of the GMM, multiple dynamic kernel-based representations are generated, which will efficiently represent the given video. The next subsections detailed the formulation of the dynamic kernels.

3.3 Dynamic Kernels

The selection of kernel function plays important role in the performance of kernel methods. For static patterns, several kernel functions are designed in past decades The kernels designed for the varying length patterns are known as dynamic kernels. Dynamic kernels are either formed by converting variable length patterns to static patterns or by designing new kernel functions. In this sub-section, we present different dynamic kernels, which effectively preserve local and global information, respectively, for better representation of the given sample.

3.3.1 Explicit Mapping based Dynamic Kernel

In the explicit mapping dynamic kernel, the set of variable length local feature representations are mapped onto fixed dimensional feature representation in the kernel space by GMM based likelihood. The Fisher kernel (FK) used for the proposed approach maps the set of variable length local features onto the fixed dimensional Fisher score. The Fisher score is computed by evaluating the first derivative of log-likelihood for mean, covariance, and weight vector using Equation 2 given by

$$\psi_q^{(\mu)}(\mathbf{V}) = \sum_{n=1}^{N} p(q|\mathbf{v}_n)\mathbf{m}_{nq}, \tag{3}$$

$$\psi_q^{(\sigma)}(\mathbf{V}) = \frac{1}{2}\left(\sum_{n=1}^{N} p(q|\mathbf{v}_n)\left[-\mathbf{u}_q + \mathbf{h}_{nq}\right]\right), \tag{4}$$

$$\psi_q^{(w)}(\mathbf{V}) = \sum_{n=1}^{N} p(q|\mathbf{v}_n)\left[\frac{1}{w_q} - \frac{p(q_1|\mathbf{v}_n)}{w_1 p(q|\mathbf{v}_n)}\right]. \tag{5}$$

where $\mathbf{m}_{nq} = \sum_q^{-1}(\mathbf{v}_n - \mu_q)$, $\mathbf{u}_q = \Sigma_q^{-1}$ and $\mathbf{h}_{nq} = \left[m_{n1q}\mathbf{m}_{nq}^T, m_{n2q}\mathbf{m}_{nq}^T, \cdots, m_{ndq}\mathbf{m}_{nq}^T\right]$. For any $d \times d$ matrix $\mathbf{A}$ with $a_{ij}, i, j = 1, 2, \cdots, d$ as its elements, $vec(\mathbf{A}) = [a_{11}, a_{12}, \cdots, a_{dd}]^T$.

The first-order derivative or the gradient of the log-likelihood computed above represent the directions in which the parameters, namely, μ, Σ, and w should be updated for the best fit of the model. We infer that the deviations that occurred, during the facial movements of particular expressions are captured by these gradients. The fixed dimensional feature vector known as the Fisher score vector is then computed by stacking all the gradients from Equation 3, 4, and 5 given by

$$\Phi_q(\mathbf{V}) = \left[\psi_q^{(\mu)}(\mathbf{V})^T, \psi_q^{(\sigma)}(\mathbf{V})^T, \psi_q^{(w)}(\mathbf{V})^T\right]^T. \tag{6}$$

The Fisher score vector for all the Q components of the GMM is given by

$$\Phi_s(\mathbf{V}) = \left[\Phi_1(\mathbf{V})^T \Phi_2(\mathbf{V})^T \Phi_Q(\mathbf{V})^T\right]^T. \tag{7}$$

The Fisher score vector captures the similarities across two samples, thus the kernel function for comparing two samples V_x and V_y, with given local features is computed by

$$\mathbf{K}(\mathbf{V}_x, \mathbf{V}_y) = \Phi_s(\mathbf{V}_x)^T \mathbf{F}^{-1} \Phi_s(\mathbf{V}_y), \tag{8}$$

Where $\mathbf{I}$ is knows as Fisher information matrix given by

$$\mathbf{I} = \frac{1}{D}\sum_{d=1}^{D} \Phi_s(\mathbf{V}_d)\Phi_s(\mathbf{V}_d)^T. \tag{9}$$

The Fisher information matrix captures the variability's in the facial movement across the two samples. Thus both local and global information is captured using Fisher score and Fisher information matrix in Fisher kernel computation. However, the computation complexity for the Fisher kernel is highly intensive. The computation of gradient for mean, covariance, and weight matrix involves $Q \times (N_p + N_r)$, each. Then the computation of the Fisher information matrix involves $D \times d_s^2 + D$ computations, where D is the total number of training examples. Similarly, the Fisher score vector requires $d_s^2 + d_s$ computations, where d_s is the dimension of the Fisher score vector. Thus, the total computation complexity of the Fisher kernel is given as $\mathcal{O}(QN + Dd_s^2 + D + d_s^2 + d_s)$ as shown in Table 1.

3.3.2 Probability based Dynamic Kernel

In probability-based dynamic kernels, the set of variable length local feature representations are mapped onto fixed dimensional feature representation in the kernel space by comparing the probability distributions of the local feature vectors. Initially, the *maximum aposteriori* (MAP) adaptation of means and covariances of GMM for each clip is given by

$$\mu_q(\mathbf{V}) = \alpha \mathbf{F}_q(\mathbf{V}) + (1-\alpha)\mu_q. \tag{10a}$$

and

$$\sigma_q(\mathbf{V}) = \alpha \mathbf{S}_q(\mathbf{V}) + (1-\alpha)\sigma_q. \tag{10b}$$

where $\mathbf{F}_q(\mathbf{V})$ is the first-order and $\mathbf{S}_c(\mathbf{V})$ is the second-order Baum-Welch statistics for a clip $\mathbf{V}$, respectively, which is calculated as

$$\mathbf{F}_q(\mathbf{V}) = \frac{1}{n_q(\mathbf{V})}\sum_{n=1}^{N} p(q|\mathbf{v}_n)\mathbf{v}_n \tag{11a}$$

and

$$\mathbf{S}_q(\mathbf{V}) = diag\left(\sum_{n=1}^{N} p(q|\mathbf{v}_n)\mathbf{v}_n\mathbf{v}_n^T\right), \tag{11b}$$

respectively.

The adapted mean and covariance from each GMM component depend on the posterior probabilities of the GMM given for each sample. Therefore, if the posterior probability is high then higher will be the correlations among the facial movements captured in the GMM components. This shows that the adapted mean and covariance for each GMM mixture will have a higher impact than the full GMM model means and covariances. Thus, the adapted means from Equation 10a, for sample V is given by

$$\psi_q(\mathbf{V}) = \left[\sqrt{w_q}\sigma_q^{-\frac{1}{2}}\mu_q(\mathbf{V})\right]^T. \tag{12}$$

Table 1: Statistics of the collected database score-wise in 3-score grading scales.

Kernels	Number of computations		Computational Complexity
Fisher Kernel (FK)	Gradient vector computation	$3 \times Q \times (N_p + N_r)$	$\mathcal{O}(QN + Dd_s^2 + D + d_s^2 + d_s)$
	Fisher information matrix	$D \times d_s^2 + D$	
	Kernel computation	$d_s^2 + d_s$	
Intermediate matching kernel (IMK)	Posterior probability computation	$Q \times (N_p + N_r)$	$\mathcal{O}(QN)$
	Comparisons to select features	$Q \times (N_p + N_r)$	
	Base kernel Computation	Q	
GMM supervector kernel (GMM-SVK)	Mean adaptation	$Q \times (N_p + N_r)$	$\mathcal{O}(QN + Qd_l^2 + d_s^2)$
	Supervector computation	$Q \times (d_l^2 + 1)$	
	Kernel computation	d_s^2	
GMM mean interval kernel (GMM-MIK)	Mean adaptation	$Q \times (N_p + N_r)$	$\mathcal{O}(QN + Qd_l^2 + Qd_l + Q^2 d_s^2)$
	Covariance adaptation	$Q \times (N_p + N_r)$	
	Supervector computation	$Q \times (d_l^2 + d_l)$	
	Kernel computation	d_s^2	

By stacking the GMM vector for each component, a $Qd \times 1$ dimensional supervector is obtained, which is known as GMM supervector (GMM-SV) represented as $\mathbf{S}_{svk}(\mathbf{V}) = [\psi_1(\mathbf{V})^T, \psi_2(\mathbf{V})^T, \cdots, \psi_Q(\mathbf{V})^T]^T$.

The GMM-SV used for comparing the similarity across two samples, namely, $\mathbf{V}_x$ and $\mathbf{V}_y$ by constructing GMM supervector kernel (GMM-SVK), which is given by

$$K_{svk}(\mathbf{V}_x, \mathbf{V}_y) = \mathbf{S}_{svk}(\mathbf{V}_x)^T \mathbf{S}_{svk}(\mathbf{V}_y). \quad (13)$$

The GMM-SVK kernel formed above only utilizes the first-order adaptations of the samples for each GMM components. Thus, the second-order statistics, i.e., covariance adaptations is also involved in constructing fixed-length representation from variable-length patterns is given by

$$\psi_q(\mathbf{V}) = \left(\frac{\sigma_q(\mathbf{V}) - \sigma_q}{2} \right)^{-\frac{1}{2}} \left(\mu_q(\mathbf{V}) - \mu_q \right). \quad (14)$$

Combining the GMM mean interval supervector (GMM-GMI) for each component is computed as $\mathbf{S}_{mik}(\mathbf{V}) = [\psi_1(\mathbf{V})^T, \psi_2(\mathbf{V})^T, \cdots, \psi_Q(\mathbf{V})^T]^T$.

Thus, to compare the similarity across the two samples $\mathbf{V}_x$ and $\mathbf{V}_y$, the kernel formation is performed using GMM-GMI kernel also known as GMM mean interval kernel (GMM-MIK) given by

$$K_{mik}(\mathbf{V}_x, \mathbf{V}_y) = \mathbf{S}_{mik}(\mathbf{V}_x)^T \mathbf{S}_{mik}(\mathbf{V}_y). \quad (15)$$

The fixed-length representation formed by using the posterior probabilities in the kernel space is a high dimensional vector, which involves $Q \times (N_p + N_r)$ computations for mean adaptation and $2 \times Q \times (N_p + N_r)$ for mean and covariance adaptations, respectively. And the kernel computation required $Q \times (d_l^2 + 1)$ and d_s^2, where d_l is the dimension of local feature vector. The total computational complexities of GMM-SVK and GMM-MIK kernels are $\mathcal{O}(QN + Qd_l^2 + d_s^2)$ and $\mathcal{O}(QN + Qd_l^2 + Qd_l + Q^2 d_s^2)$, respectively as shown in Table 1.

3.3.3 Matching based Dynamic Kernel

The kernels mentioned above are mentioned based on the mapping of variable-length feature representa-

Figure 2: Illustration of the facial paralysis patients posing 10 different expressions under expert supervision. Black patches are imposed to hid the identity of the patient (best viewed in color).

tions to fixed-length feature representations. This section introduces the alternative approach for designing of the new kernel for handling variable-length data, known as matching based dynamic kernels. Various matching based dynamic kernels are proposed in the literature like summation kernel (SK), matching kernel (MK), etc. However, these kernels are either computationally intensive or not proved to be the Mercer's kernel. So, an intermediate matching kernel (IMK) is formulated by matching a set of local feature vectors by closest virtual feature vectors obtained using the training data of all classes. Let $\mathbf{Z} = \{\mathbf{z}_1, \mathbf{z}_2, \cdots, \mathbf{z}_Q\}$ be the virtual feature vectors. Then, the feature vectors $\mathbf{v}^*_{xq}$ and $\mathbf{v}^*_{yq}$ in $\mathbf{V}_x$ and $\mathbf{V}_y$, respectively, that are nearest to q^{th} virtual feature vector $\mathbf{z}_q$ is determined as

$$\mathbf{v}^*_{xq} = \arg\min_{\mathbf{v}\in\mathbf{V}_x} \mathcal{D}(\mathbf{v}, \mathbf{z}_q) \text{ and } \mathbf{v}^*_{yq} = \arg\min_{\mathbf{v}\in\mathbf{V}_y} \mathcal{D}(\mathbf{v}, \mathbf{z}_q), \tag{16}$$

where $\mathcal{D}(.,.)$ is a distance function, which measures the distance of a feature vector $\mathbf{V}_x$ or $\mathbf{V}_y$ to the closest feature vector in $\mathbf{Z}$. We hypothesize that the distance function aid in finding the closest facial muscle movement learned from the clip to one, which is captured by GMM components. Once the closest feature vector is selected, the base kernel will be given by

$$K_{imk}(\mathbf{V}_x, \mathbf{V}_y) = \sum_{q=1}^{Q} k(\mathbf{v}_{xq}, \mathbf{v}_{yq}). \tag{17}$$

In the proposed approach, the GMM parameters like mean, covariance, and weight are used as a set of virtual feature vectors. And, the distance or closeness measure is computed by using the posterior probability of the GMM component generating the feature described in Equation 2. Thus, the local feature vectors close to the virtual feature vector for the given q is $\mathbf{v}*_{xq}$ and $\mathbf{v}*_{yq}$ for clips $\mathbf{V}_x$ and $\mathbf{V}_y$, respectively, which is computed as

$$\mathbf{v}^*_{xq} = \arg\max_{\mathbf{v}\in\mathbf{V}_x} p(q|\mathbf{v}) \text{ and } \mathbf{v}^*_{yq} = \arg\max_{\mathbf{v}\in\mathbf{V}_y} p(q|\mathbf{v}). \tag{18}$$

The computational complexity of IMK is very low compared to other mentioned dynamic kernels defined as (i) $Q \times (N_p + N_r)$ comparisons for selection of closest feature vector, (ii) $Q \times (N_p + N_r)$ required for posterior probability computations, and (iii) Q base kernel computations. Thus the total computational complexity of IMK is given by $\mathcal{O}(QN)$ where N is the set of local feature vector as shown in Table 1.

For classification, support vector machine (SVM) is built for each dynamic kernel. The SVM is a two-class classifier, For D training samples can be represented as $(V_d, y_d)_{d=1}^{D}$, where y_d represents the label information of the particular class, then discriminant function for SVM is given by,

$$f(V) = \sum_{d=1}^{D} \alpha^*_d \, y_d \, K_{DK}(V, V_d) + b^* \tag{19}$$

where D_s be the number of support vectors, α^* is the optimal values of the Lagrangian coefficient and b^* is the optimal bias. The sign value of the function f decides the class of V. We use a one-against rest approach with 10 fold cross-validation to discriminate the sample of the particular class with all the other classes.

4 EXPERIMENTAL RESULTS

In this section, we describes about the facial paralysis dataset in detail. Also, we analyse different types of dynamic kernels representations for better quantitative assessments. We compare the proposed approach with existing state of the art approaches and in last we discuss the efficacy of the proposed approach with some ablations study.

4.1 Dataset Collection and Annotation Protocol

To show the efficacy of the proposed approach we collected the video dataset of the facially paralyzed

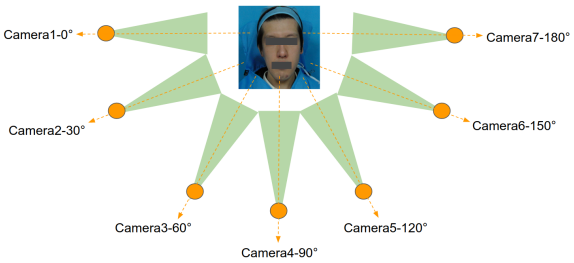


Figure 3: Camera position during the video recording of the facially paralyzed patients, black patches are added to hide the identity of the patient (best viewed in color).

patients under 3 expert supervision. The patients concerned are taken in advance for the collection of videos. Multiple subjects of various age group, gender, races, etc, are collected. Also, the video recorded is captured from seven different angle views by placing multiple cameras at different angle setting of +/- 30°as shown in Figure 3. The main objective of collecting subject and view-invariant videos of the patients is to develop an accurate and generalized model for the quantitative assessment of facial paralysis. The total number of video samples collected for the experiments is 2717 from 39 subjects. These 39 subjects are of different age starting from 17 years to 70 years, the detailed statistics of the dataset age-wise and gender-wise is shown in Figure 4. During capturing the patient videos, patients are asked to perform the 10 expressions given on Figure 2 and also subjective assessments using Yanagihara grading scale under 3 experts supervision are computed for ground truth evaluation. The experts also grade the expressions posed by the patients from score-0 to score-5. As already mention, the grading provided by the experts are highly subjective, thus, for the ground truth of the proposed model, we took 2 best subjective expert opinions out of 3 experts. Based on the subjective assessments we divided the whole dataset into 2166 training videos and 551 testing videos for score-0 to score-5 as shown in Table 2 and for score-0 to score-2 as shown in Table 3. Also, the testing video subjects are not at all present in the training set in any conditions during experimentation.

Table 2: Statistics of the collected database score-wise in 5-score grading scales.

Grading scores	# of training videos	# of testing videos	# of total videos
Score 0	166	62	228
Score 1	322	104	426
Score 2	600	147	747
Score 3	539	140	679
Score 4	539	98	637
Total videos	2166	551	2717

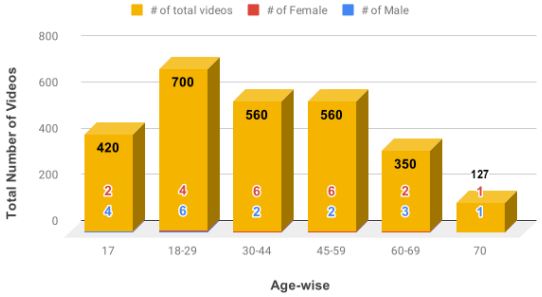


Figure 4: Statistics of the data collected age-wise and gender-wise (best viewed in color).

Table 3: Statistics of the collected database score-wise in 3-score grading scales.

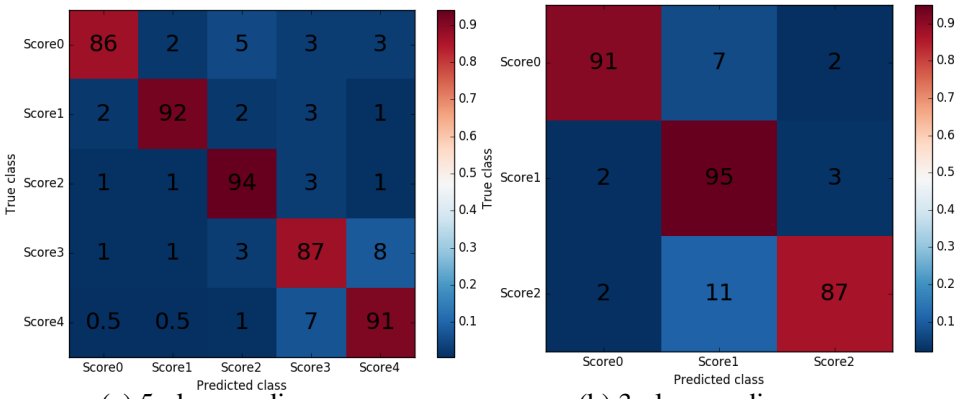
Grading scores	# of training videos	# of testing videos	# of total videos
Score 0	488	166	655
Score 1	1139	287	1426
Score 2	539	98	637
Total videos	2166	551	2717

4.2 Analysis of the Dynamic Kernels for Quantitative Assessment of Facial Paralysis

The classification performance of various dynamic kernel like Fisher kernel (FK), intermediate matching kernel (IMK), supervector kernel (SVK), and mean interval kernel (MIK) using different GMM components, namely 32, 64, 128, 256, and 512 is shown in Table 4 for 5-class grading score. The spatio-temporal facial features, namely, histogram of optical flow (HOF) and motion boundary histogram (MBH) are trained using GMM and classified using kernel-based support vector machine (SVM) Cortes and Vapnik (1995). It can be observed that the best performance kernels are probability-based kernels, namely, support vector kernel (SVK) and mean interval kernel (MIK) as it captures the first-order and second-order statistics of the learned GMM model. Also, it can be observed that increasing the number of mixtures in GMM increases the better generalization capability of the model, however, it cannot be increased beyond 256 due to increase in demand of the local feature information, which cannot be addresses due to the lim-

Table 4: Classification performance (%) of FK, IMK, SVK, and MIK on different GMM components for 5-class grading score.

# of components	FK		IMK		SVK		MIK	
	HOF	MBH	HOF	MBH	HOF	MBH	HOF	MBH
32	37.3	40.1	67.1	70.5	68.8	74.1	70.5	75.2
64	43.6	44.5	72.3	73	69.7	76.2	72	75.8
128	45.5	45.5	74.1	75.8	71.4	77.3	73	78.6
256	**47.9**	**48.6**	**76.6**	**77.9**	**78.4**	**82.2**	**86.5**	**90.7**
512	46.8	47.9	76.2	76.2	72.3	78.4	81.5	87.1



(a) 5-class grading score (b) 3-class grading score

Figure 5: Confusion matrix of MBH feature vector using GMM-MIK dynamic kernel with SVM for 256 components (best viewed in color).

ited size of the dataset.

The confusion matrix for 5-class grading score is given in Figure 5 (a), it can be observed that the misclassified samples are mostly present in the neighboring class, due to which we combined the score-0 class examples with score-1 class examples and the score-2 class examples with score-3 class examples. Following the previous work Ngo et al. (2016), NGO et al. (2016), and Wachtman et al. (2002), we reduce the number of classes from 5-class grading score to 3-class grading to facilitate the comparison of the proposed work with the state of the art approaches. The classification performance of the above fusion i.e. for 3-class grading scores are shown in Table 5 and confusion matrices for the best performances are shown in Figure 6.

Table 5: Classification performance (%) of FK, IMK, SVK, and MIK on different GMM components for 3-class grading score.

# of components	FK		IMK		SVK		MIK	
	HOF	MBH	HOF	MBH	HOF	MBH	HOF	MBH
32	52.6	53.8	68.8	71.4	71.4	72.3	85.9	87.2
64	53.7	55.2	70.5	75.8	76.8	79.2	86.9	89
128	55.2	58.2	73.2	76.2	78.8	79.9	88.9	90.8
256	**62.3**	**63.2**	**78.4**	**81.5**	**82.4**	**84.1**	**90.2**	**92.5**
512	55.4	59.9	75.8	78.6	80.2	81.7	89.6	91.5

4.3 Expression-wise Classification Performance and Comparison with the State of Art Approaches

The performance comparison with the state of the art methods is given in Table 6. Also, to show the efficacy of the proposed approach we evaluate the proposed approach with most most popular, 3DCNN features Tran et al. (2014) and classified the same using SVM.

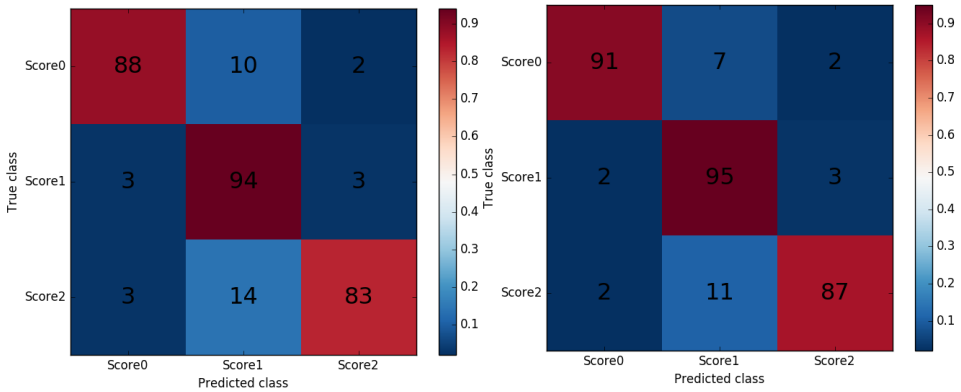
Table 6: Comparison with state of the art methods.

Methods	Accuracy (%)
PI Wachtman et al. (2002)	46.55
LBP He et al. (2009)	47.27
Gabor Ngo et al. (2014)	55.12
Tracking 2D NGO et al. (2016)	64.85
Tracking 3D Ngo et al. (2016)	66.47
C3d (from fc-8 layer and on 5-class grading scores) Tran et al. (2014) + SVM	71.5
C3d features(from fc-8 layer and on 3-class grading scores) Tran et al. (2014) + SVM	81.3
Proposed approach (on 5-class grading scores)	90.7
Proposed approach (on 3-class grading scores)	92.46

Table 7: Expression wise classification performance (%) of the proposed approach for the best model (MBH features using MIK kernel for 512 components.

	EP0	EP1	EP2	EP3	EP4	EP5	EP6	EP7	EP8	EP9
Proposed 5-score grading score	75.45	95.23	86.4	91.94	94.13	95.97	91.57	93.77	89.01	93.04
Proposed 3-score grading score	81.68	94.87	93.04	97.43	92.3	95.6	91.94	95.23	87.54	95.23

It can be observed that the proposed approach has better representative features than 3DCNN features. Also, the expression wise classification performance of the best model i.e. MBH features with MIK kernel for 256 components is given in Table 8. It can be observed that the expression with fewer facial movements like at rest expression (EP0) has lower performance as compared to the expression with prominent facial movements like the closure of eye tightly (EP3), wrinkle nose (EP5), etc. We also compare the previous works and the proposed approach expression wise in Table 9. However, it can be noticed that only a few



(a) HOF (b) MBH

Figure 6: Confusion matrix of HOF and MBH feature vector using GMM-IMK dynamic kernel with SVM for 256 components for 3-class grading scores (best viewed in color).

expressions from the previous works are compared, this is due to the previous works only focus on the expressions which have notable (eminent/distinguished) facial movements like wrinkle forehead (EP1), closure of eye tightly (EP3), wrinkle nose (EP5), and grin (EP7). This is evaluated to facilitate the comparison with the previous work.

Table 8: Expression wise classification performance (%) of the proposed approach for the best model (MBH features using MIK kernel for 256 components.

Expression Denotations	Proposed 5-score grading score	Proposed 3-score grading score
EP0	75.45	81.68
EP1	95.23	94.87
EP2	86.44	93.04
EP3	91.94	97.43
EP4	94.13	92.3
EP5	95.97	95.6
EP6	91.57	91.94
EP7	93.77	95.23
EP8	89.01	87.54
EP9	93.04	95.23

Table 9: Comparison of the classification performance (%) for the few prominent facial paralysis expressions with the existing works.

	PI	LBP	Gabor	Tracking 2D	Tracking 3D	Proposed 5-class grading scores	Proposed 3-class grading scores
EP1	50.7	58.3	62.4	69.4	70.9	95.23	94.87
EP3	48.2	48.9	53.1	62.1	63.3	91.94	97.43
EP5	48.1	41.8	50.5	57.3	58.2	95.97	95.6
EP7	39.2	40.1	54.5	70.6	73.5	93.77	95.23

4.4 Efficacy of the Proposed Approach

Figure 7 shows the visualization of the kernel matrix of the best performing MBH features with a mean interval kernel (MIK) for 256 components and 3-class grading score. The lighter shade of the diagonal elements show the higher values, which represents the correctly classified elements while the off-diagonal elements in darker shade represent the lower values. Also, it can be inferred that using MIK as a distance metric there is better separability among the different levels of the facial paralysis.

Further, it can be observed from Table 8, expressions like at rest (EP0) and closure of eye lightly (EP2), where there are few or no facial movements results in low performance of the proposed approach. Also, from figures 8 (a) and 8 (b), it can be observed that expressions having common facial movements like blowing out cheeks (EP6) and whistling (EP8) are confused with each other. And expression having distinguished (uncommon) facial movements like wrinkle forehead (EP1) and wrinkle nose (EP5) are less confused with each other.

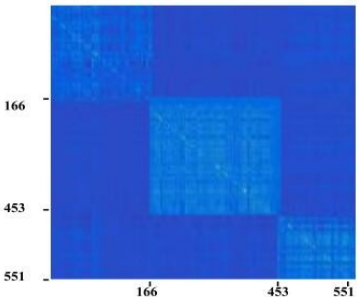


Figure 7: Mean interval kernel representation for motion boundary histogram (MBH) features and uGMM 256 components and 3-class grading score (best viewed in color).

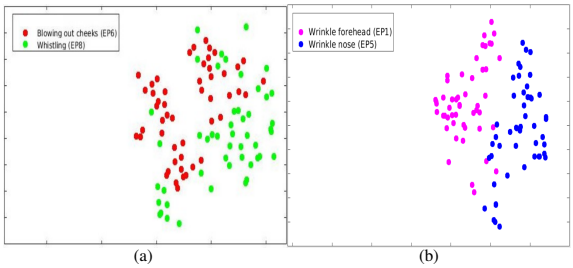


Figure 8: t-sne plot for the expressions of facial paralysis using MBH and GMM-MIK dynamic kernel based SVM for 256 components for 3-class grading score (best viewed in color). In (a) t-sne plot for expression blowing out cheeks (EP6) and whistling (EP8) and in (b) for expression wrinkle forehead (EP1) and wrinkle nose (EP5).

5 CONCLUSION

In this paper, we introduce a novel representation of the facial features for variable length pattern using dynamic kernel-based classification, which provide the quantitative assessment to the patients suffering from facial paralysis. Dynamic kernels are used for representing the varying length videos efficiently by capturing both local facial dynamics and preserving the global context. A universal Gaussian mixture model (GMM) is trained on spatio-temporal features to compute the posteriors, first-order, and second-order statistics for computing dynamic kernel-based representations. We have shown that the efficacy of the proposed approach using different dynamic kernels on the collected video dataset of facially paralyzed patients. Also, we have shown the computation complexity and classification performance of each dynamic kernels, where the matching based intermediate matching kernel (IMK) is computationally efficient as compared to other dynamic kernels. However, probability-based mean interval kernel (MIK) is more discriminative but computationally complex. In the future, the classification performance has to be improved further by improving the modeling of expressions for better quantitative assessment of the facial paralysis. Also, various quantitative assessment using Perveen et al. (2012); Perveen et al. (2018); Perveen et al. (2016) are need to be explore and compare for better classification performance.

REFERENCES

Banks, C. A., Bhama, P. K., Park, J., Hadlock, C. R., and Hadlock, T. A. (2015). Clinician-Graded Electronic Facial Paralysis Assessment: The eFACE. *Plast. Reconstr. Surg.*, 136(2):223e–230e.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine Learning*, 20(3):273–297.

Dileep, A. D. and Sekhar, C. C. (2014). 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*, 25(8):1421–1432.

Guo, Z., Shen, M., Duan, L., Zhou, Y., Xiang, J., Ding, H., Chen, S., Deussen, O., and Dan, G. (2017). Deep assessment process: Objective assessment process for unilateral peripheral facial paralysis via deep convolutional neural network. In *2017 IEEE 14th International Symposium on Biomedical Imaging (ISBI 2017)*, pages 135–138.

Hato, N., Fujiwara, T., Gyo, K., and Yanagihara, N. (2014). Yanagihara facial nerve grading system as a prognostic tool in Bell's palsy. *Otol. Neurotol.*, 35(9):1669–1672.

He, S., Soraghan, J. J., O'Reilly, B. F., and Xing, D. (2009). Quantitative analysis of facial paralysis using local binary patterns in biomedical videos. *IEEE Transactions on Biomedical Engineering*, 56(7):1864–1870.

House, J. W. and Brackmann, D. E. (1985). Facial nerve grading system. *Otolaryngology-Head and Neck Surgery*, 93(2):146–147. PMID: 3921901.

Liu, X., Dong, S., An, M., Bai, L., and Luan, J. (2015). Quantitative assessment of facial paralysis using infrared thermal imaging. In *2015 8th International Conference on Biomedical Engineering and Informatics (BMEI)*, pages 106–110.

NGO, T. H., CHEN, Y.-W., MATSUSHIRO, N., and SEO, M. (2016). Quantitative assessment of facial paralysis based on spatiotemporal features. *IEICE Transactions on Information and Systems*, E99.D(1):187–196.

Ngo, T. H., Chen, Y. W., Seo, M., Matsushiro, N., and Xiong, W. (2016). Quantitative analysis of facial paralysis based on three-dimensional features. In *2016 IEEE International Conference on Image Processing (ICIP)*, pages 1319–1323.

Ngo, T. H., Seo, M., Chen, Y.-W., and Matsushiro, N. (2014). Quantitative assessment of facial paralysis using local binary patterns and gabor filters. In *Proceedings of the Fifth Symposium on Information and Communication Technology*, SoICT '14, pages 155–161, New York, NY, USA. ACM.

Perveen, N., Gupta, S., and Verma, K. (2012). Facial expression recognition using facial characteristic points and gini index. In *2012 Students Conference on Engineering and Systems*, pages 1–6.

Perveen, N., Roy, D., and Mohan, C. K. (2018). Spontaneous expression recognition using universal attribute model. *IEEE Transactions on Image Processing*, 27(11):5575–5584.

Perveen, N., Singh, D., and Mohan, C. K. (2016). Spontaneous facial expression recognition: A part based approach. In *2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 819–824.

Satoh, Y., Kanzaki, J., and Yoshihara, S. (2000). A comparison and conversion table of 'the house-brackmann

facial nerve grading system' and 'the yanagihara grading system'. *Auris Nasus Larynx*, 27(3):207 – 212.

Tran, D., Bourdev, L. D., Fergus, R., Torresani, L., and Paluri, M. (2014). C3D: generic features for video analysis. *CoRR*, abs/1412.0767.

Wachtman, G., Liu, Y., Zhao, T., Cohn, J., Schmidt, K., Henkelmann, T., VanSwearingen, J., and Manders, E. (2002). Measurement of asymmetry in persons with facial paralysis. In *Combined Annual Conference of the Robert H. Ivy and Ohio Valley societies of Plastic and Reconstructive Surgeons*.

Wang, H., Oneata, D., Verbeek, J., and Schmid, C. (2015). A robust and efficient video representation for action recognition. *International Journal of Computer Vision*, 119(3):219–238.