Towards Automated Recognition of Facial Expressions in Animal Models

Gaddi Blumrosen[1,2*], David Hawellek[1], and Bijan Pesaran[1]
[1]Center of Neural Science, New-York University-NYU
[2]Computational Biology Center (CBC), IBM
[*]Current affiliation
Gaddi.Blumrosen@ibm.com, dh113@nyu.edu, bijan@nyu.edu

Abstract

Facial expressions play a significant role in the expression of emotional states, such as fear, surprise, and happiness in humans and other animals. The current systems for recognizing animal facial expression model in Non-human primates (NHPs) are currently limited to manual decoding of the facial muscles and observations, which is biased, time-consuming and requires a long training process and certification. The main objective of this work is to establish a computational framework for facial recognition systems for automatic recognition NHP facial expressions from standard video recordings with minimal assumptions. The suggested technology consists of: 1)a tailored facial image registration for NHPs; 2)a two-layers unsupervised clustering algorithm that forms an ordered dictionary of facial images for different facial segments; 3)extract dynamical temporal-spectral features; ,and recognize dynamic facial expressions. The feasibility of the methods was verified using video recordings of an NHP under various behavioral conditions, recognizing typical NHP facial expressions in the wild. The results were compared to three human experts, and show an agreement of more than 82%. This work is the first attempt for efficient automatic recognition of facial expressions in NHPs using minimal assumptions about the physiology of facial expressions.

1. Introduction

Facial expressions play an important role in the expression of internal emotional states in humans and other primates. Continuous recognition of primate facial expressions can, therefore, improve monitoring behavior and mental health condition [1]. For humans, facial expressions are considered universal and share many common properties across cultures [2]. Technologies for human facial emotion recognition are increasingly more automated and accurate due to enhanced computational capabilities, and the increased availability of storage [3]. Despite these advances, automatic tools to detect facial expressions and assess emotional states do not yet exist for non-human primates, hampering the development of animal models for mental health research.

Automatic Facial Expression Recognition (AFER) for humans decode set of pre-determined emotions, like happiness, sadness, anger, disgust, surprise, or fear [4]. AFER systems suffers from variability between subjects [4], and the objective difficulty in finding accurate ground truth for some emotional states such as pain [5], or depression [6]. Algorithms for AFER in humans are mostly muscle activation models based, or model-free statistical based, or on both [7]. Model-based methods, usually assume a predetermined prototypic number of expressions and are directly related to the decoding blocks of facial expressions muscle activity, as the one estimated by Action Units (AU) [8]. Each AU has its own set of muscle movements and set of facial appearance characteristics. The AUs can be used for any higher order decision making process including recognition of basic emotions. Facial Action Coding System (FACS) was built to objectively and comprehensively decode human facial expression [9], [10]. Model-free methods are based on applying statistical machine learning tools with massive training data sets with pre-labeled facial expressions, like deep learning based on convolutional neural-network [11].

For both model-based and model-free techniques, the algorithms consist of the following four stages: 1) face detection such as the Viola-Jones algorithm; 2) registration, to compensate over variations in pose, viewpoints (frontal vs. profile views), and across a wide range of illuminations, including cast shadows and specular reflections [12]; feature extraction like AUs' activation level, Gabor features [13], Histogram of Oriented Gradients (HOG) [14]; 4) and classification of instantaneous facial expression, or dynamic facial expression [15].

In animals, in particular in NHPs, facial expressions are key-source for communication, and related to facial dynamical gestures. In rhesus monkeys, facial expressions are sometimes linked to body postures [16], and calls [17]. In Chimpanzees, facial expressions can indicate internal emotional states, and thus play an important role in communication [18]. The pioneering work in [16], defined the main six principal facial expressions used by the rhesus monkey: 1) threat, which typically includes exposed teeth, a wide open mouth and narrowing of the

eyes; 2) fear grin, expressed through exposed teeth, closed mouth and eventual teeth grinding; 3) lip smacking, a pro-social gesture expressed by producing a smacking sound through repetitive lip movements; 4) chewing; 5) gnashing of teeth ; and 6) yawning. The latter three, are considered miscellaneous facial expressions, and have a weak link to emotional states.

Common practice for facial recognition in NHPs is analyzing video streams or snapshots by an expert, mostly by using published guidelines and clustering to different facial expression groups [19]. In the past few years, after the development of FACS in humans, methods used for recognizing facial expressions in NHPs have been based primarily on the model based approach, where the AUs are decoded and used as features in a classification algorithm [20]. However, applying the FACS designed for humans to NHPs is not feasible due to the differences in the muscle structure between humans and NHPs, which results in differences in the facial expressions [21]. As an example, human AU 17 (the movement of the chin boss and lower lip), is largely a forward rather than an upward action in nonhuman anthropoids. A model-based approach for NHPs yielded the coding system of ChimpFACS [20], and macFACS [22], for chimpanzee, and macaque, respectively. Coding of the AUs is performed manually on still images or video snapshots by at least one expert rating. Nevertheless, independent movement of several muscles sometimes cannot be identified in FACS, although it could be determined that they were active in collaboration with other movements. For example, lip smacking is an action that involves rapid and repeated movements of the lips. However, due to the absence of lip eversion in the Rhesus macaque, it is unclear whether this movement involves AU23-Lip Tightening or AU24-Lip Pressor, or some combination of both. Thus, a single Action Descriptor, AU18i, most exclusively associated with the action is given in the macFACS [22]. At second stage, the histogram of the AUs intensity values is used for classification based on known facial expression categories labels [20].

While many AFER for humans, few efforts have been made to in NHPs. Existing model-based methods are limited to manual decoding of the AUs, which is time-consuming and requires a long training process and certification. Manual decoding is also not fully objective, as it is affected by inter-coder variability [17]. Another limitation of model-based approaches is the difficulty in detecting all appearance characteristics of the AUs related to the facial expression, in particular where facial areas are covered by hair and can hide some of the muscle activities [20]. Another main challenge is interpreting and categorizing the different NHP facial expression to meaningful emotions and time-dependent gestures [23]. Consequently, the creation of a labeled data base of different NHP facial expression that can be used for validation of different classification algorithm is a cumbersome stage needed to enable AFER in NHPs [20]. FACs' representation requires estimating the dynamics of each muscle's activation separately over different activation times, which requires supervision learning of the appearances values with labeled FACs data. This labeled FACs do not currently exist for NHPs as in humans, and restrict the use of FAC based AFER in NHPs.

The main objective of this study is to establish a mechanism and tailor baseline computational tools to enable an objective automatic decoding of NHPs' facial expressions from a standard video recording. The methods suggested in this work, were applied to data from a set of experiments that recorded the facial expressions of a non-human primate (Macaca Mulatta) in a nearly frontal-face-view condition. The subject participated in different behavioral conditions that aimed at provoking a range of facial expressions to build a subject-specific library of the repertoire of facial expressions. The system was verified against FACs decoded test data for ground-truth by three different independent experts for fundamental lower and upper facial expression.

This paper's contributions are three-fold: 1) establishment of analysis pipeline for NHP's AFER with minimal prior-assumptions regarding the NHP muscle structure, that can be a baseline for technologies that will replace the tedious state of the art manual AU's decoding, and eliminate decoding errors of facial expression related to spontaneous multiple AUs' activation; 2) establishment of NHP's facial expression data base; 3) forming computational tools that include intuitive representation, can support artifact removal, include extraction of dynamic model and features that capture the nature of the NHP's facial expression gestures in the wild, different from humans' facial expression that mostly are characterized by their instantaneous face appearance.

2. Methods

The methods presented in this work, are designed to learning the statistical features of each NHP's facial expressions with minimal prior assumptions. First, the facial images of NHPs from a video stream are registered (detected, aligned and rescaled). Then, a two layers unsupervised clustering algorithm with artifact removal is used to form ordered Eigen Facial-Expression Image (EFI) dictionary for the individual NHP. The streams of facial areas in the registered facial images are matched to the dictionary EFIs and form a dynamic pattern of facial expression over time. Spectral-temporal features are derived from the patterns that are fed to classifier to match the facial expression based on training set or on prior knowledge. Figure 1 describes the main blocks of the algorithm.

2.1. Face Detection and Registration.

For NHP's facial detection, we used the Viola-Jones algorithm [24], trained on the NHP's static areas that include the eyes, and nose. Face tracking the KLT algorithm applied on randomly-chosen point features from the NHP face [25]. The point features minimize the eigenvalues of the tracker's matrix covariance [26] by applying a forward-background algorithm [27], followed by an affine transportation of the points [28]. Since the facial images from the KLT tracker, might suffer from accumulative drift over time [29], to reduce the alignment drift, and to improve the facial registration quality, we aligned the images offline to a baseline image, with a neutral expression.

2.2. Images pre-processing and dimensionality reduction

A pre-processing stage of background removal was applied to improve the robustness to variations in background due to changes in head poses, or in the background, using a CIE tri-stimulus [30] and k-means clustering to background and facial images similar to [31]. The images were transferred to monochromatic intensity images (black and white, with 8 bits representation for each image's pixel) and were resized to a constant size of $N_r \times N_c$. The m'th intensity image of the video stream after registration, background removal, and resizing, is denoted by I_c^m, $m = 1 \dots M$.

2.3. Establishing the Eigen Facial-Expressions Image (EFI) Dictionary

For the clustering algorithm, a Principle Components Analysis (PCA) was applied on the raw data images, representing the images with M vectors of size N_0. The PCs are fed to the two-layer clustering and forms an Eigen Facial-expressions Images (EFIs) set that represent typical facial expressions that are close to the ones used in the training video.

The first clustering layer transforms from the M image dimension to a lower observable EFIs features space of size N_1, where $M \gg N_1$. For this, we use the k-mean clustering that is performed on the PCA domain to save computational resources and find the mapping function L_1 that minimizes the cluster square error. These clusters can be represented by their mean image value and form a set of N_1 EFIs of size $N_r \times N_c$ each, where the i′th EFI is the mean value of the cluster's images:

$$E_I^{n_i} = E\left(\hat{L}_1^i(I_c^m)\right), \quad (1)$$

where E is the expectation operator, and $\hat{L}_1^i$, is the i′th cluster indices vector, $\hat{L}_1$.

The EFIs encode a set of different facial appearances that represent a combination of one or more facial appearances related to the activation of multiple AUs in different facial expressions [32] [20], [22]. The EFI representation, with sufficient number, represents a set of facial appearances that span of facial expressions with error decreasing with the dictionary size [33]. This enables a compact interruptible representation in compare to AU representation [34] and to other non-model based representation [35].

Artifacts in facial expression videos can be caused by alignment errors, inaccurate background removal, or instantaneous blocking of the face, for example by hand movement or by other subjects in the scene. For this, statistical tools like blind source separation can be applied similar to the one used for example in neural signals [36].

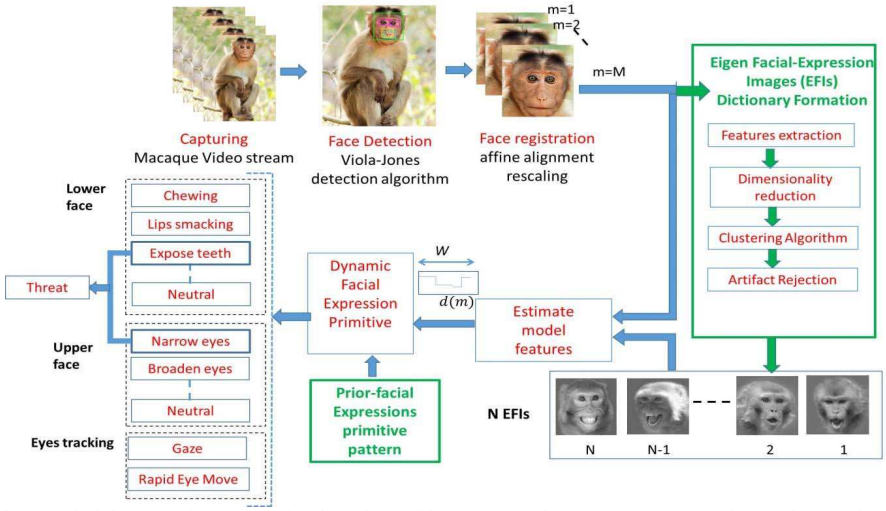

Fig. 1. The suggested non-human facial expression estimation based on video stream scheme. It composes of two phases: data collection, and real-time analysis. The first stage of data collection uses the stream of video of images to form a facial-expression image dictionary, using a blind clustering algorithm. The facial expression dictionary includes Eigen Facial-expression Images (EFI), which are sorted according to their similarity. The second stage, which can work in real time, and on multiple facial areas, finds the most likelihood facial image from the dictionary that matches the instantaneous video image frame. The facial expression are taken from the work in [23].

Automatic artifact removal methods can be also used based on deviation from "normal" distributions, based on cluster variance, or the silhouette error matrix [37]. The EFI's representation of facial appearances enables also a human expert visual inspection process for artifact removal. The number of valid EFIs after the artifact removal is equal to N_1 minus the number of artefactual EFIs'.

The EFIs from the first-layer clustering were then ordered to induce similarity between proximate EFIs, to enables sub-clustering, and to reduce estimation error in EFI estimation [34]. The ordering is performed on the clustered EFIs after performing the Gabor transform, to increase sensitivity to local directionality of the different facial parts [13]:

$$L_2 = sort(G(E_I)), \quad \text{s.t. } D \tag{2}$$

where E_I is set of EFIs $\{E_I^n\}_{n=1}^{n=N_2}$, G is Gabor transform, and $sort$, is hierarchical clustering operation subject to similarity measure D, such as minimal Euclidean distance [38].

This result of the clustering is the ordered set of EFIs:

$$\mathrm{E_I'} = \left\{\mathrm{E_I^{n'}}\right\}_{\mathrm{n'=1}}^{\mathrm{n'=N_2}}, \tag{3}$$

where the indices are given by $\mathrm{n'} = \mathrm{L_2(n_i)}, \mathrm{n'} = 1 \ldots \mathrm{N_2}$.

2.4. Extension to sub-facial areas analysis

The muscle activations can be independent across facial areas when muscle activity is not coordinated or acts in different phases. An example for non-coordinated muscle activity is the "Brow Raiser" and "Lips towards each other" (AU8). This implies that facial expression clustering can be improved by working independently on different facial areas. The fundamental separation between the lower and upper part of the face was shown to be informative in facial expression recognition [39]. Some other facial regions such as ears or eyes (measured by tracking the pupil location) can be also decoded separately [20]. Similar to human facial expression recognition, a commonly used boundary between the lower and upper facial areas, based on AUs' correlation and distribution, is the Infra orbital furrow (an area near the nostrils) [32],[39]. The lower part includes facial morphological features (landmarks) of the chin, mouth corners, philtral region, sub nasal furrow, and the nasal groove. The upper facial part include the brow, eyelid furrow, glabella, and the cheeks [22]. For the lower part, there are 14 identified AUs (related to lips, jaw, mouth, cheeks), while for the upper part 4 main identified AUs ("Brow raiser", "Glabella lowerer", "lid raiser", "Lid tightener") [22]. For the ears there are three active AUs (Ears forward, elevator, and flattener).

Without loss of generality, we look on lower and upper facial areas after artifact removal and the two-level clustering, which performed separately on each area:

$$\mathrm{E_{I_L}'} = \left\{\mathrm{E_{I_L}^{n_L'}}\right\}_{\mathrm{n_L'=1}}^{\mathrm{n_L'=N_L}}, \text{ and } \mathrm{E_{I_U}'} = \left\{\mathrm{E_{I_U}^{n_U'}}\right\}_{\mathrm{n_U'=1}}^{\mathrm{n_U'=N_U}}. \tag{4}$$

where $\mathrm{N_L}$, and $\mathrm{N_U}$, are the number of lower and upper EFIs' set number.

Another facial area is the eyes' pupils, which is considered in many facial expression technologies like FACS (Facial Action Coding System) [9], as part of the facial expression. The low intensity color of the pupils or an Eigen-eye pattern, can be used for eye tracking [40]. In this work, we use the eye movements' locations relative to the eye center as a feature to facial expressions recognition. Since the eye movements from both sides are usually coordinated and correlated [41], we can look only on the mean or standard deviation of the two eyes' displacements. Quantizing the displacement can further reduce this feature space dimension. Let's denote $\hat{\mathrm{d}}_\mathrm{h}$, $\hat{\mathrm{d}}_\mathrm{v}$, as the horizontal and vertical planes estimations of the eye tracker like the one in [40].

For the representation of diverse facial appearances in each facial region, the number of EFIs in each facial area (size of dictionary) should be higher than the total AUs' appearances combinations in this area. In case the muscles' activation (AUs) are independent, the number of facial appearances in each area would have been multiplication of all possible AUs' appearances. This number can be very high. Since the AUs in each facial area are usually coordinated, this number decreases with the level of correlation between the AUs in the region.

2.5. Spatial-temporal Features Extraction

The sorted EFIs in (4), and (5) were matched to the stream of images $\mathrm{I_c^m}$ and formed model stream of images with the highest similarity to the instantaneous facial image according to:

$$\hat{n}^{FR} = argmin_{n^{FR}} \left(I_{FR}^m - E_{FR}^{n^{FR}}\right)^2, \tag{5}$$

where I_{FR}^m, is the m'th facial region image, and $\hat{n}^{FR}$, is the estimated dictionary index for the facial region FR.

The estimated EFIs' indices can be concatenated and form a temporal waveform that contains spatial-temporal information that can be used for recognition of dynamic changes in facial expression:

$$d^{FR}(k) = \{\hat{n}^{FR}(m)\}_{m=(k-1)W+1}^{m=kW}, \tag{6}$$

where the window length W, should be tuned to capture the dynamic across facial variations, possibly throughout the all expression cycle [42].

The features can be the indices histogram, spectral features, like the frequencies with maximal response, mean and standard deviation.

2.6. Dynamic Facial Expressions Recognition

While humans facial expressions, are related mostly to

emotional states [16], the primitives of NHPs facial expressions in NHPs, are characterized more by their dynamic characteristics. A Maximum Likelihood (ML) estimator can match to each facial area separately, to recognize facial expression primitives:

$$\hat{F}_{ep}^{I} = argmax_{F_{ep}^{I}} P\left(D^{I}|F_{ep}^{I}\right), \quad (7)$$

where I, is indicator for the facial area, $I = \in \{L, U, E\}$, F_{ep}^{L}, F_{ep}^{U}, and F_{ep}^{E}, and D^{L}, D^{U} and D^{E}, are the facial areas primitives, and feature set, for the lower, higher facial areas, and eyes, respectively.

Lower Facial expression primitives can be exposing teeth, or chewing (which is related to many muscles activations around the mouth), lip smacking (which is more related to periodical facial expression changes) and neutral. Upper facial expression primitives can be opening and narrowing the eyes, frowning, or eyebrow rise. Eye related primitives can be gaze, staring, or moving the eye rapidly (rapid eye movement).

Together, the recognized primitives can be input to second layer classifier to form more abstract facial expression, that can be more correlated to internal emotional state like feat or threat [16].

3. Experimental Setup

The main goal of this study was to demonstrate the new methodology and show its capability to recognize NHPs facial expressions. The experimental setup included a NHP subject, and a standard video recording camera.

The NHP subject was a male rhesus macaque (monkey M, 6 kg) that trained to be seated in a chair and be head-fixed such that head rotations were limited. The fixation however enabled the NHP to freely eat, and to produce facial expressions typical for situations of social interactions. All animal care procedures were approved by the New York University Animal Care and Use Committee and were performed in accordance with the National Institute of Health guidelines for care and use of laboratory animals.

Two sessions of around 12 minutes were recorded with a video camera (LifeCam Microsoft Inc.) at a frame rate of 15 Hz, resulting in two data streams of length 11986, and 11395. The camera was placed at a location that did not interfere with the subject's line of sight and yet could obtain a sufficiently high resolution of the face. Each session was composed of 4 sub-sessions of around 3 minutes long, and each under different social conditions designed to obtain a large range of facial expressions. The conditions and their durations for the two sets were: 1) the subject alone (0-3:50, 0-3:20); 2), in front of a mirror (3:51-7:20, 3:21-6:24); 3) being fed (7:21-10:20, 6:25-9:20); and 4) in visual range of another conspecific, to enable possible communication (10:21-14:05, 9:21-12:30), respectively. For the last condition another NHP, likewise chaired, was brought into the same room at a distance of about six feet and the pair was free to engage in visual and communicative interaction.

A dedicated SW written in Matlab(R) (2016a, Matlab Inc,) that implemented the algorithm described in section 2 was used. To evaluate the classification performance 76 short lower facial expression video recording clips with duration of 2-7 seconds were derived from the video recordings. The facial expression classes were: neutral (20), lip smacking (10), chewing (24), and random mouth opening (22). Three independent unbiased experts in NHP behavior (MR, BF, and NB) rated the clips following [16], and [22]. More details related the expert facial decoding are provided in the Appendix. The image areas of the ears, that capture ear movement, were not included in this analysis. For eye blinking, a visual inspection (with running the video in playback in quarter of its original speed) was used as "ground truth" to evaluate eye blink detection.

4. Results and Discussion

For the initial face detection, the Viola-Jones algorithm [24] was used as described in 2.1. The eyes and nose were detected and were used as a base for forming a rectangular facial area tolerant to minor deformation of the face, like mouth opening. An inner part of the face was used for KLT tracking (maximal bidirectional Error of 2, three Pyramid Levels, and block size of 31×31. Figure 2a. shows the subject as captured by the video lens without processing. The black rectangular is the detected facial area from the Viola-Jones algorithm, and the internal yellow one, is the area for KLT tracker, where the green crosses are the point features capturing facial edges. An alignment procedure relative to a reference neutral facial expression image (mouth closed, eye opened) was performed using an arbitrary set of 1300 points. Each image was resized to 180×145 pixels, then the background was removed as shown in Figure 2.b. Figure 2.c, shows the lower, upper facial areas (130×145, and 50×145 pixels), and the green shows the eye pupils facial areas.

A PCA was performed on the registered images concatenation of the two sets. K-mean clustering was applied with 50 clusters. This number was chosen to be greater than the number of AUs in the facial region, reflecting average correlation of the AUs of around 0.3 (6), and compromises minimal dimensionality that captures most of the facial expressions in the training set, and a sufficient number to capture diversity. The set of 50 mean values of the clusters form the EFIs according to (2). Clustering quality was estimated by the silhouette diagram [37], and was 0.25, with 95% of the 50 clusters having average positive value, which indicates on relatively separated clusters and supported the chosen number of clusters. Then artefactual clusters, for example, due to blockage of the facial image by hand, were excluded from

dictionary based on deviation of their statistics from the main facial clusters, similar to [36].

Then, the EFIs the lower, and upper facial areas' EFIs, with $N_L = 30$, and $N_U = 20$, were derived. The EFIs of the different areas were sorted using hierarchical clustering on the images after applying Gabor transform (filter size of 39×39, number of scales of 5, and number of orientations of 8). Figure 3.a, and Fig. 3.b, describes the EFIs E'_{I_L}, and E'_{I_U}, after sorting, where the number in brackets indicate on the EFIs before sorting. For the lower EFIs (E'_{I_L}), the first six EFIs, can be associated to different levels of mouth opening. The following 22 EFIs seems to be more around the neutral image, which indicate that they decode more settles AUs' activities. For the upper EFIs, it seems that EFIs 1-3, have different levels of narrowing the eyes, while the others decode settle muscle activity of the upper part.

The EFIs indices streams for the facial areas were estimated using the criterion in (5). The EFIs' sorting, induces a relation of proximity between consecutive indices, and enable reducing estimation noise by smoothing of the indices scores. Figure 4 shows three typical results of choosing EFIs' indices from the second experiment set. Figure 4.a, shows a neutral facial expression at time 0:01 seconds, where the EFIs, are typically chosen from near the center of the dictionary histogram. Figure 4.b shows a selection of EFI that is related to partial mouth opening (EFI 25). The upper EFI remains neutral (upper EFI 13). Figure 4.c shows a selection of an EFI that is related to narrowing the eyes (upper EFI 1).

The NHPs facial expression, are characterized more than humans, by their facial muscles' dynamics. Figure 5, shows the lower area facial expression dynamics from observation of 0.6 sec (9 frames) for lip smacking, and chewing. The lower facial areas are relatively close to the neutral facial expression. Figure 5.a2, shows the standard deviation of the sequence. Slight displacement facial areas of the lower nose and mainly the areas around the lips, mainly between the upper lip and the nose can be seen as the brighter areas in the facial image. These changes are characterized by small displacement, and deformation (wrinkle), and correspond to lips related AUs (AU18 Lip pucker, AU10 Upper lip raiser, and AU8 Lips toward each other), which are associated with lip smacking, and of AU9 Nose wrinkle. For Chewing, Fig 5.b2, the nose wrinkle, and lips related muscles still exists, but lower lip (AU16 Lower lip depressor), and mainly jaw related muscle (AU26, Jaw drop, and AU27, Mouth stretch) are active.

Figure 6 shows the EFIs indices' estimation (EFI streams) for the two experiment-sets' repetitions. In this study, we choose 66 states to decode all facial expressions, 50 for the facial areas, and 16 for the horizontal and vertical eye pupils' locations. The lower EFIs in Fig. 6.a,

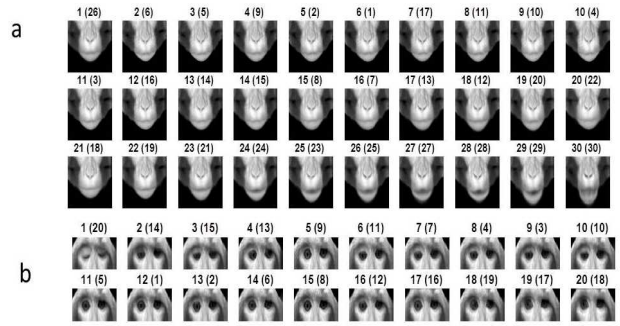

Fig. 3. (a), and (b), describes the EFIs E'_{I_L}, and E'_{I_U}, after sorting, where the number in brackets indicate on the EFIs before sorting.

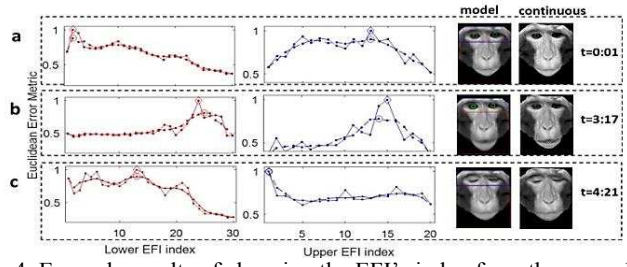

Fig. 4. Example results of choosing the EFI's index from the second experiment set. (a): a neutral facial expression; (b) partial mouth opening; (c) narrowing the eyes. The line values represent the likelihood of the image to be modeled by the EFI, and the dashed line is filtered version, that can be applied due to the hierarchical order of the EFIs. The marker is the maximal value of the curve, which represent the approximated EFI number.

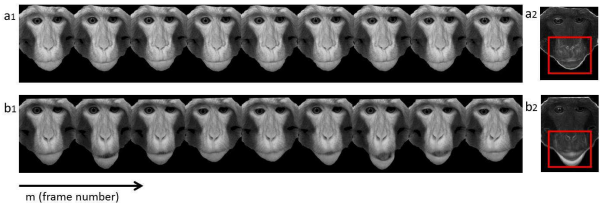

Fig. 5. Dynamics of Lip smacking (panel a), and chewing (panel b). Panels a_1, and b_1, describe the video frames of the facial expression, and panel a_2, and b_2, the standard deviation of the sequences.

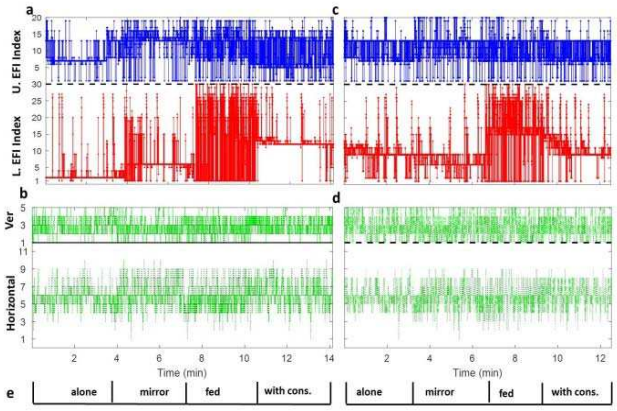

Fig. 6. EFIs' stream for the two recording sessions. Panels (a) and (c) are related to the first one, and (b) and (d), to the second. Panel (e), shows the different experiment conditions. The blue, red, and green colors represent the upper, lower, and eyes region.

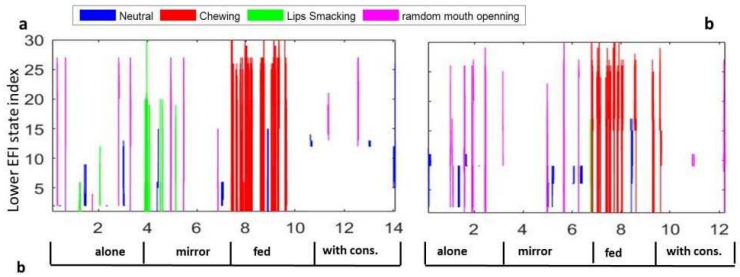

Fig. 7. (a): The 76 labeled video clips of the Lower EFI states for the two recording sessions.

and 6.b, in the time where the NHP was left alone as shown in panel 6.e, seems to be flat with relatively low variability, and having EFI represent close mouth. At the second stimulation (mirror), there are high amplitude fluctuations that include the full and partial mouth opening (EFIs 23 to 30) for both repetitions. The third phase of eating seems to be determined by fluctuations of the EFIs in almost all the EFIs' range. The fourth phase seems to be similar to the first phase. The similarities between the repetitions are correlated to similar behavioral response that is captured by the suggested method. From the upper EFIs blinking patterns can be extracted by examining when the upper indices moves between neutral and close eyes EFIs (EFIs 1-3). From the pupils' movement shown in Fig. 6c, and 6.d, the low horizontal eyes position variability can indicate on a state of gaze.

The feasibility of the analysis scheme is demonstrated on recognizing lower facial expression primitives. For this, a data base of 76 short video clips was created for four fundamental lower facial expression primitives of neutral, chewing, lip smacking, or other mouth opening like one related to yawning, or teeth exposure. The EFI's indices pattern in a window length of the video clip duration of around 2 seconds, were used to extract the features. Figure 7.a, and Fig. 7.b, shows the histogram of the lower EFIs, and their distribution. Neutral facial expression is characterized by lower EFIs indices. Chewing is characterized by concentration of the EFIs in the upper region related to mouth opening and of jaw muscle activation. The lower EFIs in the chewing are possibly related to times of short break in chewing, or at start or end of the action, and can be seen as artifact; Lip smacking has neutral and mid-range EFIs' indices, which can be explained by moving periodically from neutral state to state were muscles like the lips are contracted (related to AU18i in macFACS [22]). The other lower facial expression primitive of open mouth or teeth exposure, are distributed around wide EFI's range, and reflect the activation of multiple lower AUs (AU9+10+AU12+AU16). The recording sessions shown in Fig. 7.b, have similar distribution up to a small shift in the distribution of the second experiment. From the distribution, Neutral and Chewing are more distinguishable, while Lip smacking and Random-mouth opening distributions have higher overlap.

The EFIs' indices' mean, and standard deviation, were used as features. The Neutral features are separated from other facial expressions with low mean value due to their related low EFI indices. Lip smacking has a small mean value but higher standard deviation, due to the periodic nature of muscle contractions. The Chewing and Random mouth opening features are less separated in this plane, as both involved many muscle activations, which result in similar variance distribution. To capture the dynamic nature of the facial expression, and exploit the periodical nature of facial expressions like lip smacking, spectral features of lower EFI's median frequency and its related median amplitude were derived and shown in Fig. 7.d. The amplitude of the chewing is the highest, and well separated from the other facial expression features. The frequency of the lip smacking is concentrated around 4 Hz. But the Neutral higher frequency values than expected, which can be explained by imperfect alignment. The random mouth opening has frequencies values higher than 4 Hz, which can be explained by non-periodic muscle movements with high frequency content.

A SVM classifier with 5-fold cross-validation was applied using the EFI's mean, standard deviation and the spectral median and peak amplitude features above. The confusion matrix presented in Fig. 8.a shows that the neutral and chewing are very separated from each other (100% classification success). Lip smacking was recognized erroneously as neutral for 20% of the times, and as random mouth opening for another 20% of the time. Misclassification can be explained by low amplitude lip smacking amplitude, or by one that varies in time. Higher frame rate, with richer spectral information, can contribute to separate these two facial expressions better. The average true positive rate was 81.9% (for 94 segments). Figure 8.b, shows the significancy of the features. The spectral amplitude feature is more significant then the median frequency and the statistical features have around similar significancy.

5. Conclusion and future work

In this paper we have established a mechanism and coarse computational tools to enable the objective decoding of NHP facial expressions from a standard video recording using minimal prior assumptions. The suggested methods have minimal assumptions about anatomical muscle structure, unlike FACs-based methods that are

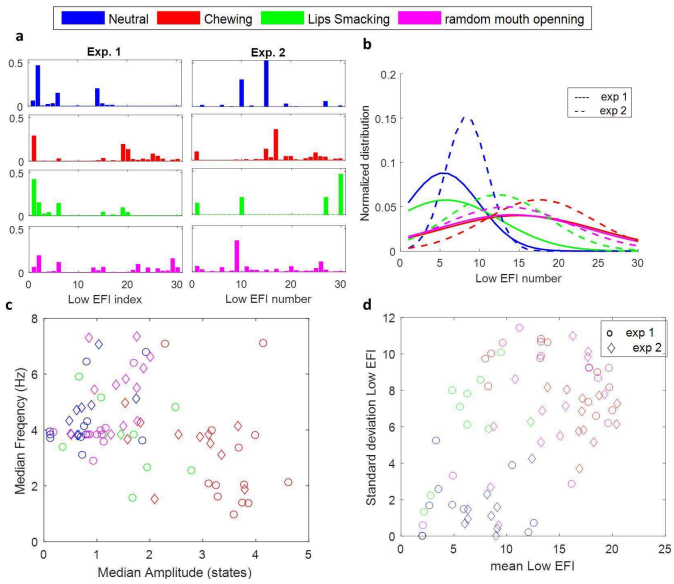

Fig. 7. Panels (a), and (b), show the EFIs' features; Panel (c), shows the histogram of the lower EFIs. Panel (d), shows the EFI's distribution. The two symbols represent the two recording sessions, which show consistency in the feature representation.

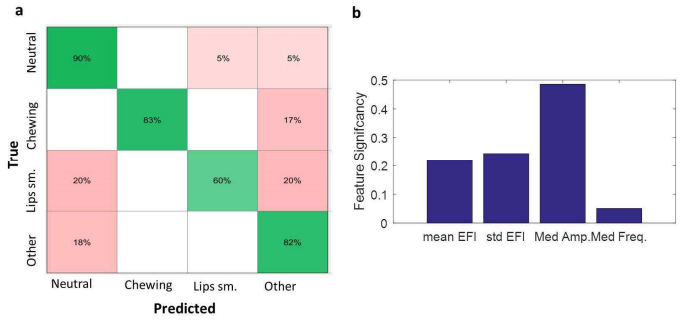

Fig. 8. (a), and (b), shows the EFIs' features; (c) shows classification results based on the four features; (d) shows the Feature significancy using F-test.

currently the state of the art methods in NHP's facial recognition and typically require an expert for decoding NHP's facial expression. The proposed representation enables detection and rejection of artifact, which makes them well-suited to being used to tracking animals' behavior in the wild. The sorting of the dictionary words, induce proximity between the EFIs, and enable extraction of informative spectral features, which are essential in recognition NHP's facial expressions like lip smacking. The set of primitive facial expression from different facial areas can be later combined to classify internal state of the NHP. We validated the recognition performance against unbiased and independent expert tagging. Facial expression recognition of Lip smacking, Teeth exposure, Neutral, and Eye movements has reached an accuracy of around 82%, with only 4 fundamental features derived from the matching EFI indices in each video clip.

In future, as the data based will increase, more features, direct estimation of the NHP's AUs, and deep learning based methods can be deployed. The similarity in facial expression between different population from the same species, and between humans, should be investigated. This is a long effort that requires massive collection and labeling to form an adequate training set. Aggregation of the facial expression estimations with other behavioral measures like NHP's body movements or voice, can also be topic for future research.

6. Acknowledgments

We would like to thank, Marsela Rubiano, Breonna Ferrentino, and Nia Boles for their efforts in tagging the lower facial expression video clips, Eshkol Fund Mr. Avraham and Mrs. Rivka Blumrosen encouragement (GB), Leopoldina Fellowship Programme Grant (LPDS/LPDR 2012-09) (DH), and an Award from the Simons Collaboration on the Global Brain (BP).

7. Appendix

Data is available upon request. The facial expressions video clips were tagged by three independent unbiased experts in NHP behavior following [16], and [22], with a majority of voting protocol.

References

[1] G. C. Littlewort, M. S. Bartlett, and K. Lee, "Automatic coding of facial expressions displayed during posed and genuine pain," Image Vis. Comput., vol. 27, no. 12, pp. 1797–1803, 2009.

[2] P. Ekman, "Facial expressions," Handb. Cogn. Emot., vol. 16, pp. 301–320, 1999.

[3] D. S. Bishwas Mishra, Steven L. Fernandes, Abhishek K , Aishwarya Alva, Chaithra Shetty, Chandan V Ajila and P. S. Harshitha Rao, "Facial Expression Recognition Using Feature based techniques and Model based techniques: A Survey 1," 2ND Int. Conf. Electron. Commun. Syst. 2015) Proc., no. Icecs, pp. 589–594, 2015.

[4] P. Ekman, W. V Friesen, and P. Ellsworth, Emotion in the human face: Guidelines for research and an integration of findings. Elsevier, 2013.

[5] M. S. Bartlett, G. C. Littlewort, M. G. Frank, and K. Lee, "Automatic decoding of facial movements reveals deceptive pain expressions," Curr. Biol., vol. 24, no. 7, pp. 738–743, 2014.

[6] J. Cohn, "Advances in Behavioral Science Using Automated Facial Image Analysis and Synthesis [Social Sciences]," Signal Process. Mag. IEEE, vol. 27, no. 6, pp. 128–133, 2010.

[7] E. Sariyanidi, H. Gunes, and A. Cavallaro, "Automatic Analysis of Facial Affect : A Survey of Registration , Representation , and Recognition," vol. 37, no. 6, pp. 1113–1133, 2015.

[8] M.-H. Yang, D. J. Kriegman, and N. Ahuja, "Detecting faces in images: A survey," Pattern Anal. Mach. Intell. IEEE Trans., vol. 24, no. 1, pp. 34–58, 2002.

[9] T. Ying-Li, T. Kanade, and J. F. Cohn, "Recognizing action units for facial expression analysis," Pattern Anal. Mach. Intell. IEEE Trans., vol. 23, no. 2, pp. 97–115, 2001.

[10] G. Littlewort, T. Wu, J. Whitehill, I. Fasel, J. Movellan, and M. . Bartlett, "CERT Computer Expression Recognition Tool."

[11] S. Lawrence, C. L. Giles, a C. Tsoi, and a D. Back, "Face recognition: a convolutional neural-network approach.," IEEE Trans. Neural Netw., vol. 8, no. 1, pp. 98–113, 1997.

[12] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," IEEE Trans. Pattern Anal. Mach. Intell., vol. 25, no. 9, pp. 1063–1074, 2003.

[13] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition.," IEEE Trans. Image Process., vol. 11, no. 4, pp. 467–476, 2002.

[14] O. Déniz, G. Bueno, J. Salido, and F. De La Torre, "Face recognition using Histograms of Oriented Gradients," Pattern Recognit. Lett., vol. 32, no. 12, pp. 1598–1603, 2011.

[15] G. Zhao and M. Pietikainen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 29, no. 6, pp. 915–928, 2007.

[16] R. a Hinde and T. E. Rowell, "Communication by Postures and Facial Expressions in the Rhesus Monkey," Proc. Zool. Soc. London, vol. 138, pp. 1–21, 1962.

[17] M. Davila-Ross, G. Jesus, J. Osborne, and K. Bard, "Chimpanzees (Pan troglodytes) produce the same types of 'laugh faces' when they emit laughter and when they are silent," PloS one, 2015.

[18] L. a. Parr and B. M. Waller, "Understanding chimpanzee facial expression: insights into the evolution of communication.," Soc. Cogn. Affect. Neurosci., vol. 1, no. 3, pp. 221–228, Dec. 2006.

[19] L. A. Parr, M. Cohen, and F. de Waal, "Influence of Social Context on the Use of Blended and Graded Facial Displays in Chimpanzees," Int. J. Primatol., vol. 26, no. 1, pp. 73–103, Feb. 2005.

[20] L. a Parr, B. M. Waller, S. J. Vick, and K. a Bard, "Classifying chimpanzee facial expressions using muscle action.," Emotion, vol. 7, no. 1, pp. 172–181, 2007.

[21] S. D. Dobson, "Allometry of facial mobility in anthropoid primates: Implications for the evolution of facial expression," Am. J. Phys. Anthropol., vol. 138, no. 1, pp. 70–81, 2009.

[22] L. a. Parr, B. M. Waller, a. M. Burrows, K. M. Gothard, and S. J. Vick, "Brief communication: MaqFACS: A muscle-based facial movement coding system for the rhesus macaque," Am. J. Phys. Anthropol., vol. 143, no. 4, pp. 625–630, 2010.

[23] L. L. a Parr and M. Heintz, "Facial expression recognition in rhesus monkeys, Macaca mulatta," Anim. Behav., vol. 77, no. 6, pp. 1507–1513, 2009.

[24] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," Comput. Soc. Conf. Comput. Vis. Pattern Recognit., vol. 1, p. I-511-I-518, 2001.

[25] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," Proceedings of the 7th international joint conference on Artificial intelligence (IJCAI'81), vol. Volume 2. pp. 674–679, 1981.

[26] C. Tomasi, "Good Features," Image (Rochester, N.Y.), pp. 593–600, 1994.

[27] Z. Kalal, K. Mikolajczyk, and J. Matas, "Forward-backward error: Automatic detection of tracking failures," Proc. - Int. Conf. Pattern Recognit., pp. 2756–2759, 2010.

[28] C. Tomasi, "Detection and Tracking of Point Features Technical Report CMU-CS-91-132," Image Rochester NY, vol. 91, no. April, pp. 1–22, 1991.

[29] H. Wang, A. Kl??ser, C. Schmid, and C. L. Liu, "Action recognition by dense trajectories," Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit., pp. 3169–3176, 2011.

[30] H. D. Cheng, X. H. Jiang, Y. Sun, and J. Wang, "Color image segmentation: Advances and prospects," Pattern Recognit., vol. 34, no. 12, pp. 2259–2281, 2001.

[31] S. L. Phung, a. Bouzerdoum, and D. Chai, "A novel skin color model in ycbcr color space and its application to human face detection," Proceedings. Int. Conf. Image Process., vol. 1, pp. 289–292, 2002.

[32] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 10, pp. 974–989, 1999.

[33] M. Aharon, M. Elad, and A. Bruckstein, "An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," Signal Process. IEEE …, vol. 54, no. 11, pp. 4311–4322, 2006.

[34] R. Ptucha and A. Savakis, "Manifold based sparse representation for facial understanding in natural images," Image Vis. Comput., vol. 31, no. 5, pp. 365–378, 2013.

[35] I. Matthews and S. Baker, "Active Appearance Models

Revisited," Int. J. Comp. Vis., vol. 60, no. 2, pp. 135–164, 2004.
[36] K. T. Sweeney, H. Ayaz, T. E. Ward, M. Izzetoglu, S. F. McLoone, and B. Onaral, "A methodology for validating artifact removal techniques for physiological signals," IEEE Trans. Inf. Technol. Biomed., vol. 16, no. 5, pp. 918–926, 2012.
[37] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," J. Comput. Appl. Math., vol. 20, pp. 53–65, 1987.
[38] D. Lin, "Facial expression classification using PCA and hierarchical radial basis function network," J. Inf. Sci. Eng., vol. 1046, no. November 1999, pp. 1033–1046, 2006.
[39] J. N. Bassili, "Emotion recognition: the role of facial movement and the relative importance of upper and lower areas of the face.," J. Pers. Soc. Psychol., vol. 37, no. 11, pp. 2049–2058, 1979.
[40] Z. Zhu and Q. Ji, "Robust real-time eye detection and tracking under variable lighting conditions and various face orientations," Comput. Vis. Image Understanding, 2005.
[41] W. H. Zangermeister and L. Start, "Types of Gaze Movement: Variable Interactions of Eye and Head Movements," Exp. Neurol., vol. 77, no. 3, pp. 563–577, 1982.
[42] S. Chen, Y. Tian, Q. Liu, and D. N. Metaxas, "Recognizing expressions from face and body gesture by temporal normalized motion and appearance features ☆," IMAVIS, vol. 31, no. 2, pp. 175–185, 2013.