A Symmetry-based Hybrid Model to Improve Facial Expressions Prediction in the Wild During Conversational Head Movements

Arvind Kumar Bansal
Department of Computer Science
Kent State University
Kent, OH, USA
email: arvind@cs.kent.edu

Mehdi Ghayoumi
eCornell and Department of Computer Science
Cornell University and University of San Diego
San Diego, CA, USA
email: mg948@cornell.edu

***Abstract*— Human emotion prediction is an important aspect of conversational interactions in social robotics. Conversational interactions involve a combination of dialogs, facial expressions, speech modulation, pose analysis, head gestures, and hand gestures in varying lighting conditions and noisy environment involving multi-party interaction. Head motions during conversational gestures, multi-agent conversations and varying lighting conditions cause occlusion of the facial feature-points. Popular Convolution Neural Network (CNN) based predictions of facial expressions degrade significantly due to occluded feature-points during extreme head-movements during conversational gestures and multi-agent interaction in real-world scenarios. In this research, facial symmetry is exploited to reduce the loss of discriminatory feature-point information during conversational head rotations. CNN-based model is augmented with a new rotation invariant symmetry-based geometric modeling. The proposed geometric model corresponds to Facial Action Units (FAU) for facial expressions. Experimental data show hybrid model comprising a CNN-based model, and the proposed geometric model outperforms the CNN-based model by 8%-20%, depending upon the type of facial-expression, beyond partial head rotations.**

Keywords-Artificial Intelligence; conversation; deep neural network; emotion analysis; facial expression analysis; facial occlusion; facial symmetry; head movement; multimedia.

I. INTRODUCTION

Due to an aging population in the developed world and limited workforce, there is a growing need of social robotics for elderly care and healthcare [1]-[3]. To show empathy, interact, and converse with humans, social robots need to understand human emotions and pain in the wild [4]-[9].

Predicting emotions in the wild is complex and requires multimodal media analysis involving dialogs, voice-modulation (including timed silence), gestures (including postures, gaze, conversational head and hand gestures, and haptic gestures), facial expressions, pain and tears [10]-[22]. Many desirable human-robot interactions, such as conversational gestures, including human warmth and affection, frustration, irritation, encouragement, impatience and pain shown by a combination of voice-modulation, speech-phrases, gestures, facial expressions and haptic touch are yet to be achieved [13]. Compared to emotions exhibited in dialogs, utterances and gestures, facial expressions are exhibited more involuntarily and express a major subset of expressed human emotions and acute pain [5]-[7], [10]-[20].

Interpreting facial expressions is context sensitive and augmented with other modalities such as speech or scene analysis [11], [12], [14]. Facial expressions and their intensity vary by gender, age and culture. A subset of facial expressions has universally accepted interpretations, and current-day research on facial expression analysis of basic emotions and pain assumes universally accepted meanings [5]-[7], [15].

In real-life scenarios, a face continuously moves during a conversation based upon 1) conversational gestures, such as argumentation, interrogation and denial; 2) intensity of emotion; 3) multi-party interactions, changing ambient lighting and shadows with head-movements [13]. Head rotations stochastically occlude feature-points causing information loss hindering accurate facial-expression classification as illustrated in Fig. 1.

In cognitive psychology, two approaches study facial expressions: 1) valence and intensity based Plutchik's eight emotion classes and their subcategories; 2) basic six emotions (*anger*, *disgust*, *fear*, *happiness*, *sadness*, and *surprise*) popularized by Ekman and others [5], [6]. Computational recognition of facial expressions is based on analysis of facial video modeled by Facial Action Unit System (FACS) [6], [15]-[17].

Figure 1. An example of recognizing facial expression in the wild. [Image source: Wikimedia Commons, public domain; Credit: US Navy, Bureau of Medicine and Surgery, 1945; Available at: https://commons.wikimedia.org/wiki/File:Navy_nurse_signing_cast_--_WWII.jpg]

Many Facial Action Units (FAUs) are also associated with acute pain that interferes with emotion analysis [18]-[20]. In this paper, we focus on the recognition of facial expressions under the assumption that conversing agents do not suffer from pain.

Previous studies are mostly limited to the frontal facial view or statically aligned poses using curated databases showing nonoccluded facial expressions in proper lighting conditions [15], [23]-[36]. Recent augmentation of CNN-based modeling with *Long Short-Term Memory* (LSTM), transfer learning and multiple feed-forward neural networks (FNN) improve the prediction of facial expression during head movements [37]. However, the model does not handle extreme information loss beyond partial occlusion and does not exploit facial symmetry. Experiments with CNN-based model show that facial expression prediction drops by 10-20% for partial occlusion (less than 45° rotation) and by 30-50% beyond 45° rotation as described in section V.

CNN-based models need to restore occluded discriminatory feature-points for beyond the partial occlusion in conversational head-gestures, such as emotional disagreement, interrogation, argumentation or denial; multi-party interaction that involves significant occlusion of one part of the face. Luckily, even during extreme head-rotations, only one side of the face is occluded. Hence, facial symmetry can be used to reconstruct the occluded discriminatory feature-points by estimating the angle of facial rotation and knowing the coordinates of their counterparts on the nonoccluded side.

This research improves facial-expression analysis under head-motion by utilizing facial symmetry along the vertical major axis [38]-[40]. Prediction uses 1) estimation of the angle of facial rotation using nonoccluded feature-points; 2) inherent facial symmetry around the vertical axis of the face; 2) differences between the symmetrical points and the actual geometric feature-points from the previous frames.

The proposed hybrid model integrates CNN-based classification for partially occluded space and the symmetry-based geometric model classification beyond partially occluded space. The proposed symmetry-based geometric model also provides motion continuity and temporal context to the CNN classifier for selecting the nearest static alignment.

The major contributions in this research are:

1. Development of a symmetry-based geometric model corresponding to *Facial Action Units* (FAUs) to recover discriminative feature-points during conversational head-rotations in real-time scenarios;
2. Augmentation of the CNN-based model with the proposed symmetry-based geometric model to improve the temporal context and the facial expression prediction beyond partial occlusion.

The overall roadmap is as follows. Section II describes background concepts. Section III describes the related work. Section IV describes the proposed symmetry-based geometric model. Section V describes an overall architecture. Section VI describes the implementation and discusses experimental results. Section VII discusses the limitations and concludes the paper.

II. BACKGROUND

A. Facial Muscles and FACS System Correspondence

A combination of facial muscles expressing facial expressions and pain, is shown in Fig. 2. The associated muscles and their functions are described in Table I. Italicized descriptions in Table I mark the FAUs involved in both pain and facial expressions. The compression of muscles is externally visible through facial feature-points, as illustrated in Fig. 3. A combination of movement of the feature-points forms the basis of geometric modeling and *Facial Action Unit System* (FACS).

Facial expression analysis is based on mapping a subset of FAUs to basic facial expressions. Tables II describes the FAUs associated with the simulations of the six basic facial expressions [6], [15]-[17].

B. Facial Feature-points and Symmetry

There are two types of facial feature-points: *fixed points* and *active points*. *Fixed points* act as a reference, and *active-points* move during facial-expressions, altering x and z-coordinates of feature-points [16].

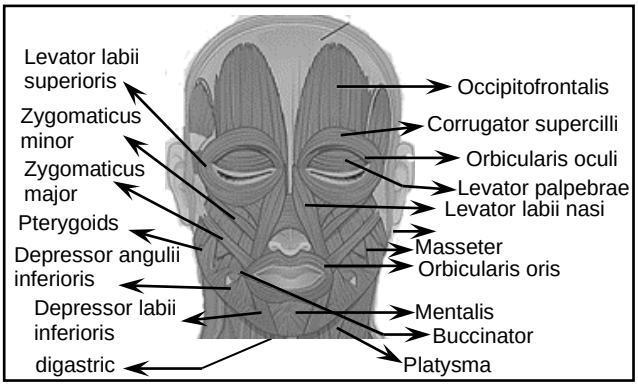

Figure 2. Facial muscles used in facial-expressions of emotion and pain. [Image adopted from Wikimedia Commons, Credit: CNX anatomy 2013]

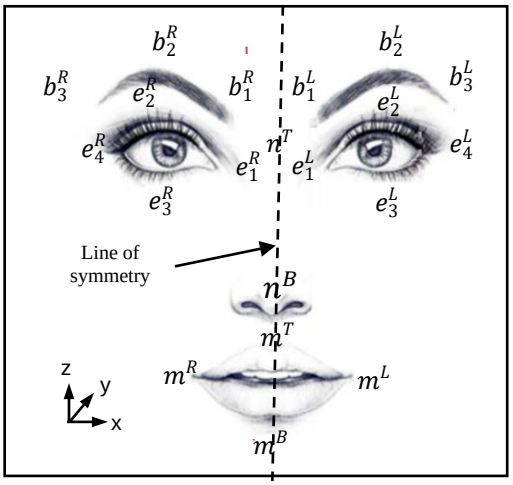

Figure 3. Facial feature points with symmetry

TABLE I. RELATED FAUS AND ASSOCIATED FACIAL MUSCLES. 'E' DENOTES EMOTION, AND 'P' DENOTES PAIN. THERE IS AN OVERLAP BETWEEN TWO SUBSETS.

FAU	Function	Facial Muscle	Type
1	inner brow raiser	occipitofrontalis	E
2	outer brow raiser	occipitofrontalis	E
4	*brow lowerer*	*corrugator supercilli*	E + P
5	Upper eyelid raiser	levator palpebrae superioris	E
6	*Cheek raiser*	*Orbicularis oculi, pars orbitali*	E + P
7	*Lid tightener*	*orbicularis oculii*	E + P
8	Lips towards each other	orbicularis oris	E
9	*Nose wrinkler*	*levator labii superioris nasi*	E + P
10	*Upper lip raiser*	*levator labii superioris*	E + P
11	Nasolabial deepener	zygomaticus minor	E
12	Lip corner puller	zygomaticus major	E
14	Dimpler	buccinator	E
15	Lip corner depressor	depressor anguli inferioris	E
16	Lower lip depressor	depressor labii inferioris	E
17	Chin raiser	mentalis	E
20	*Lip stretcher*	*platysma*	E + P
23	Lip tightener	orbicularis orbis	E
25	Lips part	depressor labii inferioris	P
26	*Jaw drop*	*masseter*	E + P
27	*Mouth stretcher*	*pterygoids, digastric*	E + P
41	*Lid droop*	*Relaxation of levator palpebrae superioris*	E + P
43	Eyes Closed	relaxation of levator palpebrae superioris	P

TABLE II. SIX BASIC FACIAL EXPRESSIONS AND FAUS

Basic facial expressions	FAU subset
Surprise	1, 2, 5, 10, 16, 26
Fear	1, 2, 4, 5, 15, 20, 26
Disgust	2, 4, 9, 15, 17
Anger	2, 4, 5, 7, 9
Happiness	6, 12, 14, 20, 27
Sadness	4, 8, 11, 15, 23, 41

Feature-point denotations use 'e' for eye; 'b' for brow; 'm' for mouth. A subscript enumerates feature-points for the same organ. A superscript denotes left-side by 'L'; right-side by 'R'; top by 'T'; bottom by 'B'.

A face has six major *fixed points* and 14 major active points. Fixed points are two ends of the left and right eyes (e_1^L, e_4^L, e_1^R, e_4^R); bottom of a nose (n^B); middle point between two eye-brows above the nose-tip (n^T). Active major points are: 1) three points on each brow (b_1^L, b_2^L, b_3^L, b_1^R, b_2^R, b_3^R); 2) two middle points of lips (m^T and m^B); 3) two endpoints of the mouth (m^R, m^L); 4) two middle points in each eye (e_2^L, e_3^L, e_2^R, e_3^R).

Facial features are symmetrical around the vertical axis as illustrated in Fig. 3: left-side of the vertical line $n^T n^B$ is symmetrical to the corresponding feature-points on the right-hand side. This symmetry causes similar changes on both sides of a face for most facial-expressions at the muscle level.

C. Occlusion and Head Movement

In a real-world situation, the head rotations are observed every 5 - 7 degrees [41]. The angle of rotation maps to one of the internal states based upon an identifiable resolution in the feature-points. Distances between the symmetry-axis and the feature-points on the nonoccluded side are used to estimate the coordinates of occluded feature-points using facial-symmetry.

D. Deep Learning Models for Facial Expression Analysis

Convolution neural network comprises a *cascade of convolution-layers*: *convolution filter*, *Rectified Linear Unit* (*RELU*) and *subsampler* (*pooling layer*) followed by a fully connected feed-forward neural network (FNN) [10], [37], [42]. In order to model a continuously moving head, multiple FNNs, one for each desired orientation, are used for the classification [37]. Due to the large number of angles, the classification is approximate, lowering the accuracy.

Long Short Term Memory (LSTM) is a variation of *Recurrent Neural Network* (RNN) where the previous outputs are fed back and gated to regulate the output [43], [44]. LSTM significantly reduces the problem of vanishing gradient and instability in RNN, and improves long range contextual dependency [44].

Transfer learning adapts meta-level learning rules to a similar but somewhat different domain or task [45]. A *domain* is a pair (ψ, $P(X)$) where ψ is the feature-space, $P(X)$ is the marginal probability-distribution, and X is a feature-vector (x_1, …, x_N) $\in \psi$. A *task* is modeled as a pair (Y, f), where Y is a label space $\{y_1, \ldots, y_M\}$, and f is a function that maps a feature-value x_i in X to a label $y_j \in Y$. The differences could be in feature-vector, probability distribution, label space, or mapping of feature-value to label space.

For facial-expression recognition, feature-vectors change due to angular variations. Thus, the feature-vector has to be approximated and modified based upon proximity and similarity analysis with input feature-vector values of an FNN to reduce classification-error.

E. Notations

Line-segments are denoted by two end feature-points or their intuitive description. For example, eye-width is denoted as EW or $e_1^L e_4^L$. Lip-width is denoted by LW or $m^T m^B$. Given a line-segment LS, magnitudes of the x-axis, y-axis and z-axis component are denoted respectively by $|LS|_X$, $|LS|_Y$, and $|LS|_Z$. In this paper, parameterization is illustrated using left-side of a face. The technique applies also to the right-side of the face.

III. RELATED WORK

In recent years, many researchers have suggested techniques to handle information loss caused by partial occlusion and multiple orientations due to head movements. Related work can be classified as: 1) handling occlusion for improper lighting conditions, hand-gestures and external objects such as eye-glasses, hats, scarfs, and medical masks; hand gestures; hair and mustaches; ambient lighting conditions; 2) analyzing emotions in the wild; 3) mapping motion as a set of fixed alignments; 4) a combination of CNN,

LSTM and transfer learning to map continuous motion to the corresponding CNN [9], [15], [23]-[31], [36], [37], [46].

Most occlusion handling schemes handle occlusion of static faces in frontal pose based on the reconstruction of small patches of partially occluded parts using the corresponding nonoccluded (or global) facial texture [23]-[31]. The schemes combine information from previous images, texture-patterns, or other features from nonoccluded space, and dynamic weighting of texture-patterns to reconstruct occluded patches using well curated datasets.

To handle the occlusion caused by external objects, researchers have trained hybrid models on occluded and nonoccluded samples and used nonoccluded features-space as a guidance to predict texture of occluded patches, combination of sparse representation and maximum likelihood estimation, a combination of Gabor filter and co-occurrence matrix of local features, a combination of Gabor descriptors and deep structure recognition based upon deep belief network, a combination of local binary pattern, feature histogram, dimension reduction and support vector machine (SVM), a fusion of information derived using Weber Local Descriptors and histograms of multiple features using SVM, combining global and local textures with CNN and attention, the use of LSTM auto-encoders and transfer learning, and Bayesian networks. These schemes do not analyze occluded facial expressions in the wild during conversational head-motion.

Zong et al. use a combination of transduction transfer learning and linear discriminant analysis to map the trained data using curated dataset to the data in the wild [9]. However, the scheme does not: 1) handle conversational head movements and the resulting occlusion; 2) does not use symmetry to recover occluded feature-points.

T-H. S. Li et al. integrate CNN with LSTM to provide the temporal context required for analyzing facial expression during head rotation [37]. They use transfer learning to map a position to the corresponding static alignment of CNN for improved accuracy. The scheme is limited by the number of fixed domains for transfer learning and does not exploit symmetry. Besides, LSTM cannot estimate the coordinates of the occluded feature-points explicitly.

These techniques are not suited for extreme loss of discriminatory feature-points during extreme head-rotations in argumentation, denial or multi-party interactions where a significant part of face is occluded for a longer period.

The proposed geometric model exploits facial symmetry to recover occluded feature-points during extreme head rotations [38]-[40]. The correspondence of the line-segments joining discriminatory feature-points to *Facial Action Units* (FAUs) relates the proposed hybrid model with *Facial Action Coding System* (FACS) based analysis and CNN-based analysis. In addition, the changes in line-segment ratios with head-movements provide temporal context even beyond partial occlusion. In our scheme, the discriminative feature-points also support multimodal analysis of head-gestures and provide explanation capability.

IV. PROPOSED SYMMETRY-BASED GEOMETRIC MODEL

Facial expression analysis requires: 1) removal of the distortions caused by camera zooming; 2) removal of the distortions in the line-segments caused by head-rotations, and 3) correspondences of parameters to the changes in FAUs.

The identification of parameters invariant to head-rotations requires the use of *fixed feature-points* that act as a reference to measure the changes in orientation and lengths of the line-segments with varying facial expressions.

The motions of *active-points* that contribute to the facial expressions are: 1) vertical and horizontal motion of b_1^L, b_2^L, b_2^L on an eyebrow; 2) vertical motions of $\{e_2^L, e_3^L\}$ in the center of an eyelid, 3) vertical and horizontal motions of m^L (lip-endpoints), and 4) vertical motions of m^T, m^B and $\{m_1^L, m_2^L\}$ (lip-midpoints). Fig. 4 shows left side of the face with the required feature-points and line-segments used in the facial expression classification.

The line-segments for the facial expression analysis are: $n^B b_1^L$, $n^B b_2^L$, $n^B b_3^L$, $b_1^L b_3^L$, *EH* ($e_2^L e_3^L$: eye-height), *LH* ($m^T m^B$: lip-height), *LW* ($m_1^L m^L$: lip width), *EL* ($m^L e_c^L$: lip segment to the eye (e_c^L is the left center of an eye given by $\frac{e_2^l + e_3^l}{2}$). The line-segments $n^B b_1^L$, $n^B b_2^L$, $n^B b_3^L$, $b_1^L b_3^L$ and *EL* have x-magnitudes and z-magnitudes.

The line-segments *LH* and *EH* have z-magnitudes; the line-segment *LW* has x-magnitude. With no rotation and zooming, changes in the x-magnitudes and z-magnitudes of these line-segments correspond to different facial expressions. In an actual scenario, these line-segments vary with head-rotations and image scaling due to the camera-zooming. These line-segments are mapped to parameters invariant to head-rotations and camera zooming, such that the resulting parameters vary with facial expressions only.

Four line-segments, joining fixed-points, $n^B n^T$, $e_1^L e_4^L$, $n^T e_1^L$, and $n^T e_4^L$ have been used to derive parameters invariant with respect to head rotation. The effect of zooming is removed by dividing the z-magnitudes by the magnitude of the line-segment $n^B n^T$.

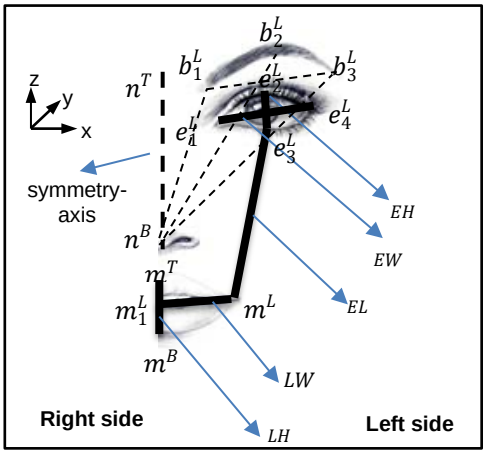

Figure 4. Facial feature-vectors

To minimize the effect of variation of x-coordinates during a head-rotation, the most *aligned fixed segments* are chosen that are affected similarly by the head-rotation compared to line-segments involving *active points*.

A division of line-segments by the x-magnitude of the line-segments involving the nearest fixed-points parallel to the same axis minimizes the effect of rotation and preserves the changes due to facial expressions.

The division by the segment $e_1^L e_4^L$ provides invariance for the eye-brow area. The division by x-magnitude $|\, n^T e_1^L \,|_X$ cancels the effect of head-rotation on the magnitude $|\, n^B b_I^L \,|_X$. The division by the x-magnitude $|\, n^T e_4^L \,|_X$ cancels the effect of the head-rotation on the magnitude $|LW\,|_x$.

A. Frontal Pose Estimation

The fixed feature-points nose-bottom n^B, left inner-eye e_1^L and right inner-eye e_1^R are used to establish frontal pose (see Fig. 3 and Fig. 4). The ratio $|\, n^T e_1^L \,| \,/\, |\, n^T e_1^R \,| = 1$ for the frontal-pose, only altering during head-rotation. The overall estimate for the frontal pose is given by (1) where ϵ is an experimentally derived value slightly greater than zero to take care of involuntary and random head-movements.

$$1 - \epsilon \leq |\, n^B e_1^L| \,/\, |\, n^B e_1^R| \leq 1 + \epsilon \qquad (1)$$

Estimation of rotation angles is based on missing landmarks on the rotated side of the face. The landmarks n^t and n^b become invisible in the complete occlusion and are visible between partial and complete occlusion. For rotation to the left or right, the ratio changes beyond $1 \mp \epsilon$.

The proposed line-segments (LH, LW, EL, EH, $|\, b_1^L\ b_3^L \,|_x$, $|n^B b_1^L|_z$, $|n^B b_2^L|_z$, $|n^B b_3^L|_z$, $n^B n^T$, EW, $|n^T n^B|$) cover all FAUs to express six basic emotions, as described in Table II. The overall correspondence is summarized in Table III.

Variations in the line-segment *LH* reflect tightening or opening of lips and mouth, and jaw-drop. It is associated with FAU 8 (lips towards each-other), FAU 10 (upper lip-raiser), FAU 16 (lower lip-depressor), FAU 17 (chin-raiser), FAU 23 (lip-tightener), FAU 26 (jaw-drop) and FAU 27 (mouth-stretcher).

TABLE III. LINE-SEGMENTS

Line-ratio	Normalized ratio	Description
R^{LH}	$\|LH\| \,/\, \|n^B n^T\|$	lip height ratio
R^{LW}	$\|LW\|_x \,/\, \|EW\|$	lip-width ratio
R^{EL}	$\|EL\|_z \,/\, \|n^B n^T\|$	eye-to-lip ratio
R^{BW}	$\|\, b_1^L\ b_3^L \,|_x \,/\, EW$	brow-width ratio
R^{IBH}	$\|\, n^B b_1^L \,|_z / \|n^B n^T\|$	inner brow-height ratio
R^{MBH}	$\|\, n^B b_2^L \,|_z / \|n^B n^T\|$	mid-brow height ratio
R^{OBH}	$\|\, n^B b_3^L \,|_z / \|n^B n^T\|$	outer-brow height ratio
R^{EH}	$\|EH\| \,/\, \|n^B n^T\|$	eye-height ratio

Variations in the line-segment *LW* reflect compression and stretching of a mouth. It corresponds to FAUs 6, 12, 14, 20, 23 and 27. These FAUs are involved in *happiness* (lip-corner and cheek-stretching obliquely up), and *sadness* (lip-corner stretching obliquely downwards).

Variations in the z-component $|EL|_Z$ (eye-to-lip vertical component) measure compression and stretching of cheek muscles. The decrease in $|EL|_Z$ corresponds to FAU 6 (cheek-raiser) associated with *happiness*. The increase in $|EL|_Z$ corresponds to FAU 15 (lip-corner depression) associated with negative emotions *fear, disgust* and *sadness*. The change in the magnitude of the line-segments *EW* (eye-width) and *EH* (eye-height) correspond to FAU 7 associated with *anger*. The magnitude $|EH|$ increases during *anger* due to the raising of the upper eyelid and middle eye-brow point.

Variations in eye-brow length $|\, b_1^L\ b_3^L \,|$ (brow compression and stretching) correspond to FAU 1 (inner brow raiser), FAU *2* (upper brow raiser) or 4 (brow lowerer). However, only the x-component $|\, b_1^L\ b_3^L \,|_x$ is used because vertical variations in eye-brow are processed by $|n^B b_1^L|_z$, $|n^B b_2^L|_z$ and $|\, n^B b_3^L \,|_z$. The increase in $|\, b_1^L\ b_3^L \,|_x$ corresponds to FAU 4 (brow-lowerer) associated with negative emotions: *fear, disgust*, *anger*, and *sadness.*

The z-component $|n^B b_1^L|_z$ corresponds to inner-eyebrow raising or lowering. The increase in magnitude $|n^B b_1^L|_z$ corresponds to FAU 1 associated with *surprise*. The decrease in $|n^B b_{i(1\leq i\leq 3)}^L|_z$ corresponds to FAUs 4 and 9 associated with negative emotions: *fear, disgust, sadness,* and *anger*. The increase in the magnitude $|n^B b_3^L|_z$ corresponds to FAU 2 associated with *fear*.

B. Normalized Ratios

In the beginning, the frontal pose is recorded to derive the original coordinates of feature-points and the original length and orientation of line-segments. The zooming distortion and head-rotation distortions in the x-direction are removed from the feature-points and the corresponding line-segments.

Vertical segments $|n^B b_1^L|_z$, $|n^B b_2^L|_z$ $|n^B b_3^L|_z$, EH and $|EL|_Z$ are divided by $|n^B n^T|$ to derive the corresponding normalized ratios. Horizontal line-segment $|LW|_X$ and $|\, b_1^L\ b_3^L \,|_X$ are divided by $|n^T e_1^L|$ and EW, respectively. The normalized ratios are summarized in Table IV.

C. FAU Correspondence

Table V describes conditions by combining the normalized ratios across the same or different video-frames that are sampled periodically because facial expressions alter after few seconds.

The increase in the ratio R^{LH} corresponds to FAU 10 (upper lip raiser), FAU 26 (jaw-drop), and FAU 27 (mouth-stretch). The decrease in the ratio R^{LH} corresponds to FAU 8 (lips towards each other), FAU 16 (lower lip-depressor), FAU 17 (chin-raiser), and FAU 23 (lip-tightener).

The increase in the ratio R^{LW} corresponds to FAU 6 (cheek-raiser), FAU 12 (lip-corner puller), FAU 15 (lip-corner depressor), FAU 16 (lower lip-depressor), and FAU 20 (lip-stretcher). The decrease in the ratio R^{LW} corresponds to FAU 23 (lip-tightener). The increase in the ratio R^{EL} corresponds to FAU 15 (lip-corner depressor); the decrease in the ratio R^{EL} corresponds to FAU 6 (cheek-raiser). The increase in the ratio R^{EH} corresponds to FAU 5 (upper lid raiser); the decrease in the ratio R^{EH} corresponds to the FAU 7 (lid tightener) or FAU 41 (lip-stoop). The increase in the ratio R^{BW} corresponds to FAU 4 (brow-lowerer).

The increase in the ratio R^{IBR} corresponds to FAU 1 (inner eye-brow raiser); the decrease in the ratio R^{IBR} corresponds to FAU 4 (brow-lowerer). The increase in the ratio R^{OBR} corresponds to FAU 2 (outer eye-brow raiser); the decrease in R^{OBR} corresponds to FAU 4 (eye-brow lowerer).

TABLE IV. LINE-SEGMENTS AND FAU CORRESPONDENCE

Line-seg.	FAUs	Basic emotions
LH	8, 10, 16, 17, 23, 26, 27	anger, disgust, fear, sadness, surprise
LW	6, 12, 15, 16, 20, 23	happiness and sadness
EL	6, 15	disgust, fear, happiness, sadness
EH	5, 7	anger
$\lvert b_1^L\ b_3^L \rvert_x$	4	anger, disgust, fear, sadness
$\lvert n^B b_1^L \rvert_z$	1, 4, 9	anger, disgust, fear, sadness, surprise
$\lvert n^B b_2^L \rvert_z$	4, 5	fear and surprise
$\lvert n^B b_3^L \rvert_z$	2	fear
$n^B n^T$	used for vertical normalizations	
$EW, \lvert n^T n^B \rvert$	invariant with head-rotation	

TABLE V. FAUS AND NORMALIZED RATIO CONDITIONS

FAUs	Condition ($n = m + k$ and $k > 0$)
#1	$R_n^{IBR} < R_m^{IBR}$
# 2	$R_n^{OBR} > R_m^{OBR}$
#4	$R_n^{IBR} < R_m^{IBR} \wedge R_n^{MBR} < R_m^{MBR} \wedge R_n^{OBR} < R_m^{OBR}$
#5, 27	$R_n^{EH} > R_m^{EH}$
#6, 12	$R_n^{LH} < R_m^{LH} \wedge R_n^{EL} < R_m^{EL}$
#7, 41	$R_n^{EH} < R_m^{EH}$
#8	$R_n^{LH} < R_m^{LH}$
#10	$R_n^{LH} > R_m^{LH}$
#15	$R_n^{EL} > R_m^{EL} \wedge R_n^{EW} > R_m^{EW}$
#16	$R_n^{LH} < R_m^{LH} \wedge R_n^{EL} > R_m^{EL}$
#17	$R_n^{EL} < R_m^{EL}$
#20	$R_n^{LW} < R_m^{LW}$
#23	$R_n^{LW} > R_m^{LW}$
#26	$R_n^{EL} > R_m^{EL}$

A simultaneous decrease in the ratio R^{LH} and an increase in the ratio R^{EL} correspond to the activation of FAU 16 (lower-lip depressor). Simultaneous decreases in the ratios R^{LH} and R^{EL} correspond to the activations of FAU 12 (lip-corner puller) and FAU 6 (cheek-raiser). Simultaneous increases in the ratios R^{EL} and R^{EW} correspond to FAU 15 (lip-corner depression). Simultaneous decreases in the ratios R^{IBR}, R^{MBR} and R^{OBR} and increase in the ratio R^{BW} corresponds to the activation of FAU 4 (eye-brow lowerer).

V. AN OVERALL ARCHITECTURE

The proposed architecture, as illustrated in Fig. 5, comprises: 1) preprocessing and denoising module; 2) hybrid classifier module; 3) angle-based output selector module.

The *preprocessing and denoising module* denoises the data, reduces dimensions of facial images using *Locality Sensitive Hashing* (LSH) and uses Gabor filter to preserve directionality in images.

The *hybrid classifier module* comprises: 1) CNN-based classifier; 2) the proposed geometric-classifier; 3) facial orientation based output selector either from the CNN-based classifier or the proposed geometric-classifier.

There are two types of connectivity: 1) based on data flow, as shown by the solid arrows; and 2) time-stamped stream of angles emanating from the second submodule of the geometric-classifier as shown by the dashed arrows.

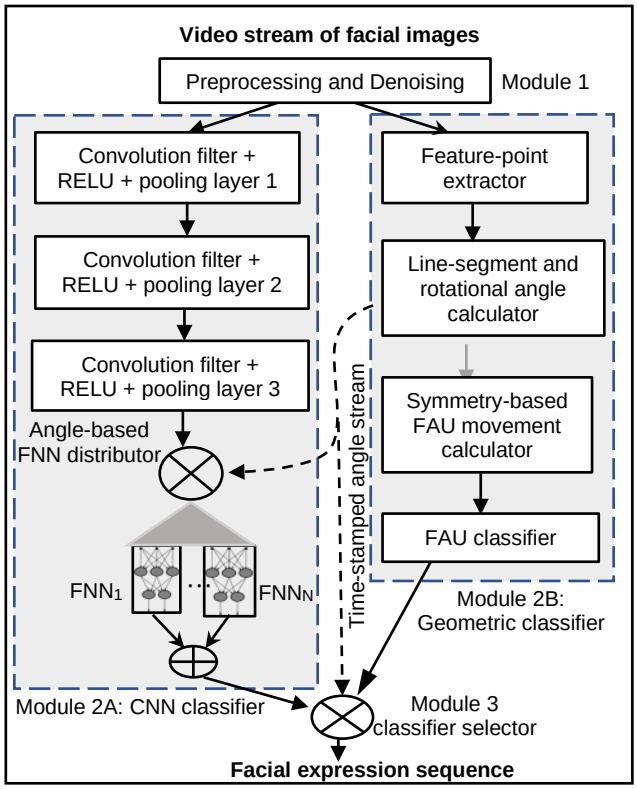

Figure 5. An overall architecture

The time-stamped angle-stream is used for: 1) selecting the output from one of the two classifiers; 2) maintaining sequentiality of the derived facial-expression labels coming from different classifiers; 3) selecting the optimal FNN in the CNN-classifier. Only one FNN is selected at a time based upon the value of the time-stamped angle.

The symmetry-based geometric-classifier has four submodules: 1) feature-point extractor; 2) invariant ratio and rotational angle calculator; 3) symmetry-based movement calculator; 4) FAU-based classifier.

Feature-point extractor extracts the visible non-occluded feature-points, calculates and time-stamps the line-segments as described in Table IV. The stream of time-stamped line-segments is passed to the second submodule.

Invariant ratio and Rotational Angle Calculator submodule derives the computable ratio $|n^B e_1^L| / |n^B e_1^R|$, $|n^T e_1^L| / |n^T e_1^R|$, $|n^T e_1^L| / |n^T n^B|$, $|n^B e_1^L| / |n^T n^B|$, $|n^T e_1^R| / |n^T n^B|$, or $|n^B e_1^R| / |n^T n^B|$, and looks up archived lookup tables to estimate the angle of rotation for each time-stamped feature-vector. This time-stamped angle is transmitted to: 1) FNN selector submodule within the CNN-classifier; 2) the third submodule of the geometric-classifier; 3) the classifier selector module. The submodule also computes the motion invariant ratio, as described in Section IV.

Symmetry-based movement calculator computes the normalized ratio conditions, as given in Table V for deriving the associated FAU movements. The input to the submodule is the motion invariant ratio and time-stamped angle derived in the submodule 2 of the classifier.

FAU-based classifier uses the derived motion of FAUs and their association with facial expressions (see Table II in Section II) to identify the facial expression [6], [16]. The input to the submodule is FAU motions derived in the third submodule of the geometric-classifier.

The CNN-classifier comprises three submodules: 1) three layers of cascaded convolution filters; 2) angle-based FNN distributor; 3) mutually exclusive FNNs that can recognize a facial-expression in one of the static orientations. FNNs are optimally trained for specific orientation in the lab conditions to improve the performance.

The choice of three layers in the convolution layer cascade and number of FNNs in the CNN-classifier is based upon size of the cropped images (56 × 56 pixels) after sampling and experimentation to reduce the computational overhead while maintaining sufficient accuracy.

Each FNN corresponds to a static orientation. Adjacent orientations are 15° apart instead of optimal 7-8° [41]. In our experiment, internal states change every 15° to reduce computational overhead. This choice slightly degrades (by 1-3%) the prediction accuracy for a tradeoff of reduced computational overhead.

Based upon each FNN corresponding to an orientation 15° apart, there are seven FNNs, one for each orientation (45° left rotation, 30° left rotation, 15° left rotation, frontal, 15° right rotation, 30° right rotation, 45° right rotation).

The angle-based distributor selects one of the FNN based upon the facial orientation. The angle is received from the second submodule of the geometric-classifier module.

The inputs to these FNNs are: output of CNN module's softmax layer distributed from the angle-based FNN distributor.

The *angle-based output selector module* selects the output from the CNN-classifier or geometric-classifier, depending upon the facial orientation. If the angle is less than the threshold value (45°), the output from the CNN-classifier is taken. For the head rotations greater than 45°, the output from the geometric-classifier is taken.

VI. IMPLEMENTATION AND EXPERIMENTATION

RaFD dataset was used for measuring the performance of the CNN-based model for various static alignments in different poses [35], [47]. Compared to other curated facial expression databases, RaFD gives comprehensive facial-expressions for 67 models (for all genders) with multiple camera angles and adjustment of lighting conditions [32]-[34]. We deployed 70% of the data for training, 15% validation, and 15% testing.

For the online video capturing, three frames per second were used for the facial expression analysis. The Epochs of 200 frames were used because the experimental data show that the accuracy of the facial expression recognition stabilizes around 200 frames. The stabilization of the 200 frame per second comes from the controlled lab conditions, and the hardware that we used for our experiments. The number will vary in noisy real environment.

A. CNN Classifier Implementation

The implemented CNN-based model is a cascade of three hidden layers: conv-32, conv-64 and conv-128, followed by a Softmax layer. Each *conv-m* layer contains *m* filters to extract different orientations. The conv-128 layer provides a subclassification of textures. After each convolution layer, there is a max-pooling layer for the subsampling of images. Each max-pool layer pools a 2 × 2 pixel macroblock.

After applying the Locality-Sensitive Hashing (LSH) and Gabor filter, the processed images are passed to the network of convolution layers through the input layer [17].

Each cropped image is scaled to 56 × 56 pixels. The data-size after the conv-32 layer is 56 × 56 × 32 pixels, and the output of first max-pooling layer after the conv-32 layer are 28 × 28 × 32 pixels. The output of the second max-pooling layer is 28 × 28 × 64 pixels. The output of the last hidden layer is 14 ×14 × 128. The output of the following max pooling layer is 7 × 7 × 128 pixels. Extracted features are concatenated by adding a fully connected layer at the end.

B. Result and Data Analysis

The hybrid model was executed in the wild. We use *recall* metrics to show accuracy of the proposed model because the recall gives the overall percentage of true positive. CNN model was also executed in wild for the frontal pose and compared against the results of RaFD dataset to derive the comparative deterioration of the *recall*.

In our statistical reporting of data, five occlusion states are used: 1) frontal face with no occlusion ($|\theta| < \epsilon$); 2) partial left-side or right-side occlusion (ϵ < rotation < 45°); 3) full left-

side or right-side occlusion (> 45°). Internal states map to one of the five states based upon interval inclusion.

C. Performance Evaluation and Discussion

The results are summarized in Tables VI, VII, and VIII. Tables VI and VIII show the recall values of CNN-model with RaFD dataset and the proposed hybrid model in wild, respectively. Table VII shows the confusion matrix for CNN model for a frontal pose in the wild.

Table VI illustrates CNN based prediction, even for a cured RaFD database, deteriorates quickly due to the unavailability of discriminatory feature-points on the occluded part of the face. The deterioration varies from 48% for sadness to 41% for happiness for complete occlusion.

Comparison of Tables VII and VIII illustrates that the accuracy of facial expression classification deteriorates in the wild even for the frontal pose: more for sadness (around 22%) and the least for disgust (around 6%). Even neutral face is labeled as sad for 10% of the time in the wild.

TABLE VI. RECALL IN THE CNN MODEL WITH RADB DATASET

	Right complete occl.	Right part occl.	Front no occl.	Left part occl.	Left complete occl.
sadness	49%	83%	97%	79%	48%
disgust	54%	81%	98%	88%	63%
anger	53%	81%	96%	87%	64%
fear	51%	86%	95%	81%	55%
surprise	57%	84%	98%	90%	53%
happiness	59%	85%	99%	92%	62%
neutral	54%	82%	95%	79%	51%

TABLE VII. CONFUSION MATRIX - CNN MODEL (FRONTAL) IN WILD

	sad. %	disg. %	ang. %	fear %	sur. %	happ. %	neutral
sadness	74.5	0.1	8.0	12.3	0.9	0.7	3.5
disgust	0.7	92.4	1.4	1.1	1.3	1.7	1.4
anger	6.4	2.3	79.3	2.5	1.6	2.4	5.5
fear	7.2	0.6	6.1	82.3	1.2	0.8	1.8
surprise	1.8	0.7	2.6	5.2	86.9	1.7	1.1
happiness	1.4	0.2	2.2	2.5	3.0	87.2	2.5
neutral	10.2	0.2	4.2	5.7	2.2	3.7	73.8

TABLE VIII. RECALL IN THE HYBRID MODEL IN WILD

	Right complete occlusion	Right part occlusion	Front no occlusion	Left part occlusion	Left complete occlusion
sadness	57%	68%	75%	69%	59%
disgust	70%	81%	92%	82%	70%
anger	73%	75%	79%	77%	76%
fear	66%	75%	82%	76%	67%
surprise	71%	74%	87%	76%	75%
happiness	75%	79%	87%	81%	77%

The reasons for this deterioration are: 1) mixing of facial muscles and feature-points for negative facial expressions, *sadness*, *fear* and *anger*, in real-time expressions; 2) variations in the intensity level of the expressed facial expressions in real-time; 3) continuous random head-motions during real-time facial-expressions causing noise; 4) uneven ambient lighting conditions with shadows obscuring feature-points; 5) randomly picking the video-frame may not correspond to the apex image corresponding to a facial-expression [46].

The facial expressions for the negative emotions: *sadness*, *fear*, and *anger* are often confused due to 1) the presence of common facial muscles; 2) the mixing of facial expressions in real-time; 3) improper temporal labeling during transition of a negative facial expression to another; 4) uncontrolled thought patterns affecting involuntary facial expressions in real-time. Another problem is that CNN is trained using fixed alignments, and a head-movement is approximated to one of the fixed poses.

Comparison of the occluded parts in Table VI and Table VIII shows that the hybrid model outperforms CNN-based prediction even for the curated RaFD dataset for beyond the partial occlusion. The improvement is 8% for sadness (minimum) to 21% for the happiness (maximum). In a multi-party interaction, where the change in the line-of-view may cause extreme occlusion, the hybrid model provides better accuracy and information.

VII. CONCLUSION AND FUTURE WORK

Head-motions during conversational gestures and multi-agent interactions cause extreme occlusion on one side of facial features. Automated feature-extracting and deep learning schemes are limited by the facial feature detections. Their performance degrades during extreme occlusion due to the nonavailability of discriminatory feature-points. Facial symmetry reconstructs the occluded discriminatory feature points. Combining CNN-based schemes with the proposed geometric modeling improves the performance in such a scenario by 8% – 21% beyond the partially occluded state.

The current scheme can be further improved by smoothening the derived facial-expression sequence and predicting the next facial-expression using Dynamic Bayesian Network (DBN), the knowledge of average duration of facial-expressions during emotional conversation, and sampling more video-frames for near-apex facial expressions [46].

In this paper, we have assumed that conversational agents are not suffering from any acute pain. As described in Section II, the facial-expressions for basic emotions and acute pain significantly overlap [7], [18]-[20], [48], [49]. Many times chronic pain is displaced and expressed as a combination of negative emotions, depression and anxiety [48], [49]. Many times, negative emotions and pain occur together and are inseparable [49]. Such cases can only be resolved by the knowledge of the situation, dialog understanding or scene analysis to build the needed context. For example, knowledge of cause of pain or video analysis of person's gesture or wound detection can provide sufficient context.

We are currently investigating the DBN on a sequence of facial-expressions to smoothen out the errors due to image frames missing the apex image for the corresponding facial expressions [46].

REFERENCES

[1] A. K. Bansal and M. Ghayoumi, "A Hybrid Model to Improve Occluded Facial Expressions Prediction in the Wild During Conversational Head Movements," The Tenth International Conference on Intelligent Systems and Applications (INTELLI 2021), July 2021, pp. 36-42, ISBN: 978-1-61208-882-2.

[2] M. I. Yenilmez, "Economic and social consequences of population aging the dilemmas and opportunities in the twenty-first century," Applied Research in Quality of Life, vol. 10, no. 4, pp. 735-752, Dec. 2015, doi: 10.1007/s11482-014-9334-2.

[3] D. H. García, P. G. Esteban, H. R. Lee, M. Romeo, E. Senft, and E. Billing, "Social Robots in Therapy and Care," 14th ACM/IEEE International Conference on Human-Robot Interaction (HRI 2019), March 2019, pp. 669-670, doi: 10.1109/HRI42470.2019.

[4] C. Peter and R. Beale (eds), "Affect and Emotion in Human-Computer Interaction: From Theory to Applications," LNCS 4868, Berlin / Heidelberg: Springer-Verlag, 2008, ISBN: 978-3-540-85098-4.

[5] R. Plutchik, "Emotion: A Psychoevolutionary Synthesis," New York, NY: Harper & Row, 1980.

[6] P. Ekman and W. V. Friesen, "Nonverbal Behavior," Communication and Social Interaction, P. F. Ostwald, (editor), New York, NY: Grune & Stratton, pp. 37- 46, 1977.

[7] K. M. Prakachin, "Assessing Pain by Facial Expression: Facial Expression as Nexus," Pain Research Management, vol. 14, no. 1, pp. 53-58, Jan./Feb. 2009, doi: 10.1155/2009/542964.

[8] F. Rothganger, "Computation in the Wild," [Online] Available from https://www.osti.gov/servlets/purl/1644432, accessed date; September 12, 2021.

[9] Y. Zong, W. Zeng, X. Huang, K. Yan, J. Yan, and T. Zhang, "Emotion recognition in the wild via sparse transductive transfer linear discriminant analysis, Journal of Multimodal Interfaces, vol. 10, pp. 163-172, June 2016, doi:10.1007/ s12193-015-0210-7.

[10] M. Ghayoumi, M. Thafar, and A. K. Bansal, "A Formal Approach for Multimodal Integration to Derive Emotions," Journal of Visual Languages and Sentient Systems, vol. 2, pp. 48-54, Oct. 2016, doi: 10.18293/DMS2016-030.

[11] C. M. Lee and S. Narayanan, "Towards Detecting Emotions in Spoken Dialogs," IEEE Transactions on Speech and Audio Processing, vol. 13, no.2, pp. 293-303, March 2005, doi: 10.1109/TSA.2004.838534.

[12] K. Han, D. Yu, and I. Tashev, "Speech Emotion Recognition Using Deep Neural Network and Extreme Learning Machine," 15th Annual Conference of the International Speech Communication Association, Sept. 2014, pp. 223-227, ISBN: 978-1-63439-435-2.

[13] A. Singh and A. K. Bansal, "Towards Synchronous Model of Non-Emotional Conversational Gesture Generation in Humanoids," Computing Conference, July 2021, Lecture Notes in Networks and Systems Series, vol. 283(1), K. Arai (editor), Cham, Switzerland: Springer, pp. 737-756, July 2021, ISBN 978-3-030-80118-2, doi: 10.1007/978-3-030-80119-9.

[14] J.-M. Fernandez-Dols, H. Wallbott, and F. Sanchez, "Emotion Category Accessibility and the Decoding of Emotion from Facial-expression and Context," Journal of Nonverbal Behavior, vol. 15, no. 2, pp. 107- 123, June 1991, doi: 10.1007/BF00998266.

[15] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial Expression Recognition: A Survey," Symmetry, vol. 11, no. 10, Article 1189, Sept. 2019, doi: 10.3390/sym11101189.

[16] M. Ghayoumi and A. K. Bansal, "An Integrated Approach for Efficient Analysis of Facial expressions," The 11th International Conference on Signal Processing and Multimedia Applications (SIGMAP 2014), Aug. 2014, pp. 211-219, ISBN: 978-989-758-046-8.

[17] M. Ghayoumi and A. K. Bansal, "Emotions in Robot using Convolutional Neural Network," International Conference on Social Robotics (ICSR 2016), Nov. 2016, Lecture Notes in Artificial Intelligence Series, Editor: A. Agah et el., , vol. LNAI 9979, Cham, Switzerland: Springer, pp. 285-295, ISBN: 978-3-319-47436-6, doi: 10.1007/978-3-319-47437-3_28.

[18] K. D. Craig, K. M. Prkachin, R. V. Grunau, "The facial expression of pain," In: D. C. Turk and R. Melzack, eds., Handbook of Pain Assessment, 3rd edition, New York: Guilford, pp. 117-133, 2011, . ISBN 978-1-60623-976-6.

[19] A. C. de C. Williams, "Facial Expression of Pain: an Evolutionary Account," Behavioral and brain sciences, vol. 25, no. 4, pp. 439–455, Aug. 2002, doi: 10.1017/S0140525X0200 0080.

[20] P. Lucey, J. F. Cohn, I. Matthews, S. Lucey, S. Sridharan, J. Howlett, K. M. Prkachin, "Automatically Detecting Pain in Video Through Facial Action Units," IEEE Transactions on System, Man and Cybernetics, vol. 41, no. 3, pp. 664-674, June 2011, doi: 10.1109/TSMCB.2010.2082525

[21] A. Gračanin, E. Krahmer, M. Balsters, Dennis Küster, and A. J. J. M. Vingerhoets, "How Weeping Influences the Perception of Facial Expressions: The Signal Value of Tears," Journal of Nonverbal Behavior, vol. 45, pp. 83–105, March 2021, doi: 10.1007/s10919-020-00347-x.

[22] N. J. Emery, "The Eyes Have it: The Neuroethology, Function and Evolution of Social Gaze," Neuroscience and Behavioral Reviews, vol. 24, issue 6, pp. 581-604, Aug. 2000, doi: 10.1016/S0149-7634(00)00025-7.

[23] L. Zhang, B. Verma, D. Tjondronegoro, and V. Chandran, "Facial Expression Analysis Under Partial Occlusion: A Survey," ACM Computing Surveys, vol. 51, no. 2, Article No. 25, April 2018, doi: 10.1145/3158369.

[24] S. S. Liu, Y. Zhang, and K. P. Liu, "Facial Expression Recognition under Random Block Occlusion Based on Maximum Likelihood Estimation Sparse Representation," International Joint Conference on Neural Networks (IJCNN 2014), July 2014, pp. 1285-1290, doi: 10.1109/IJCNN32435. 2014.

[25] L. Shuaishi, Z. Yan, and L. Keping, "Facial-expression Recognition under Partial Occlusion Based on Weber Local Descriptor Histogram and Decision Fusion," The 33rd Chinese Control Conference (CCC 2014), July 2014, pp. 4064-4068, doi: 10.1109/CCC32826.2014.

[26] Q. Cheng, B. Jiang, and K. Jia, "A Deep Structure for Facial Expression Recognition under Partial Occlusion," The Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing (IIH-MSP 2014), Kitakyushu, Japan, Aug. 2014, pp. 211-214, ISBN: 978-1-47995-391-2.

[27] J. Y. R. Corenjo and H. Pedrini, "Recognition of Occluded Facial Expressions based on CENTRIST Features," IEEE International Conference on Acoustics, Speech and Signal Processing, March 2016, pp. 1298-1302, ISBN: 978-1-47999-988-0.

[28] R. Li, P. Liu, K. Jia, and Q. Wu, "Facial Expression Recognition under Partial Occlusion Based on Gabor Filter and Gray-level Co-occurrence Matrix," The International Conference on Computational Intelligence and Communication Networks (CICN 2015), Dec. 2015, pp. 347-351, doi: 10.1109/CICN.2015.6.

[29] Y. Li, J. Zeng, S. Shan, and X. Chen, "Occlusion Aware Facial Expression Recognition Using CNN with Attention Mechanism," IEEE Transactions on Image Processing, vol. 28, no. 5, pp. 2439-2450, May 2019, doi: 10.1109/TIP.2018. 2886767.

[30] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, "Robust LSTM-Autoencoders for Face De-Occlusion in the Wild," IEEE Transactions on Image Processing, vol. 27, no. 2, pp. 778-790, Feb. 2018, doi: 10.1109/TIP.2017.2771408.

[31] Y. Miyakoshi and S. Kato, "Facial Emotion Detection Considering Partial Occlusion of Face using Bayesian Network," IEEE Symposium on Computers & Informatics (ISCI 2011), March 2011, pp. 96-101, ISBN: 978-1-61284-689-7.

[32] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A Complete Dataset for Action Unit and Emotion-specified Expression," International Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010), San Francisco, CA USA, June 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[33] R. Gross, I. Matthews, J. F. Cohn, T. Kanade, and S. Baker, "Multi-PIE," In Proceedings of the Eighth IEEE International Conference on Automatic Face and Gesture Recognition, vol. 28, no. 5, May 2008, pp. 807-813, doi: 10.1016/j.imavis.2009.08.002.

[34] M. J. Lyons, M. Kamachi, and J. Gyoba, "Japanese Female Facial-expressions (JAFFE), " 1998, doi: 10.5281/zenodo.3451524.

[35] O. Langner, R. Dotsch, G. Bijlstra, D. H. J. Wigboldus, S. T. Hawk, and A. van Knippenberg, "Presentation and Validation of the Radboud Faces Database," Cognition & Emotion, vol. 24, no. 8, pp. 1377-1388, Nov. 2010, doi: 10.1080/ 02699930903485076.

[36] K. Seshadri and M. Savvides, "Towards a Unified Framework for Pose, Expression, and Occlusion Tolerant Automatic Facial Alignment," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 38, no. 10, pp. 2110-2122, Oct. 2016, doi: 10.1109/TPAMI.2015.2505301.

[37] T-H S. Li, P-H Kuo, T-N Tsai, and P-C Luan, "CNN + LSTM Based Facial Expression Analysis Model for a Humanoid Robot," IEEE Access, vol. 7, pp. 93998-94011, July 2019, doi: 10.1109/ACCESS.2019.2928364.

[38] S. Derrode and F. Ghorbel, "Shape Analysis and Symmetry Detection in Gray-level Objects using the Analytical Fourier-Mellin Representation," Signal Processing, vol. 84, no. 1, pp. 25-39, Jan. 2004.

[39] S. Kondra, A. Petrosino, and S. Iodice, "Multi-scale Kernel Operators for Reflection and Rotation Symmetry: Further Achievements," IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2013), June 2013, pp. 217-222, doi: 10.1109/CVPRW.2013.4.

[40] M. Ghayoumi and A. K. Bansal, "Real Emotion Recognition by Detecting Symmetry Patterns with Dihedral Group," Third International Conference on Mathematics and Computers in Sciences and in Industry (MCSI 2016), Chania, Greece, Aug. 2016, pp. 178-184, doi: 10.1109/MCSI.2016.041.

[41] M. Amiri, G. Jull, and J. Bullock-Saxton, "Measuring Range of Active Cervical Rotation in a Position of Full Head Flexion using the 3D Fastrack Measurement System: an Intra-tester Reliability Study," Manual Therapy, vol. 8, no. 3, pp. 176-179, Aug. 2003, doi: 10.1016/s1356-689x(03)00009-2.

[42] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, et el., "Review of Deep Learning: Concepts, CNN Architectures, Challenges, Applications, Future Directions" Journal of Big Data, vol. 8, Article 53, 2021, 74 pages, doi: 10.1186/s40537-021-00444-8.

[43] Y. Yu, X. Si, C. Hu, J. Zhang, "A Review of Recurrent Neural Networks: LSTM Cells and Network Architectures," Neural Computation, vol. 31, pp. 1235-1270, 2019, doi: 10.1162/ neco_a_01199.

[44] K. Greff, R. K. Srivastava, J. Koutník, B. R. Steunebrink, and J. Schmidhuber, "LSTM: A Search Space Odyssey," Transactions of Neural Networks and Learning Systems, vol. 28, issue 10, pp. 2222-2232, Oct. 2017, doi: 10.1109/TNNLS. 2016.2582924.

[45] S. J. Pan and Q. Yang, "A Survey of Transfer Learning," IEEE Transactions on Knowledge and Data Engineering, vol. 20, no. 10, pp. 1345 – 1359, Oct. 2010, doi: 10.1109/ TKDE.2009.19.

[46] A. Cruz, B. Bhanu, and N. S. Thakoor, "Vision and Attention Theory-Based Sampling for Continuous Facial Emotion Recognition," IEEE Transactions of Affective Computing, vol. 5, no. 4, pp. 418-431, Oct-Dec. 2014, doi: 10.1109/TAFFC.2014.2316151.

[47] A. R. Dores, F. Barbosa, C. Queirós, I. P. Carvalho, and M. D. Griffiths, "Recognizing Emotions Through Facial Expressions- A Large Scale Experimental Study," International Journal of Environmental Research and Public Health, vol. 17, Article 7420, Oct. 2020, doi: 10.3390/ ijerph17207420.

[48] Z. Chen, R. Ansari, and D. Wilkie, "Automated Pain Detection from Facial Expressions using FACS: A Review," [Online], arXiv:1811.07988v1, Available at https://arxiv.org/pdf/ 1811.07988.pdf .

[49] M. S. H. Aung, S. Kaltwang, B. Romera-Paredes, B. Martinez, A. Singh, M. Cella, et al., "The Automatic Detection of Chronic Pain-Related Expression: Requirements, Challenges and the Multimodal EmoPain Dataset," IEEE Transactions of Affective Computing, vol. 7, no. 4, pp. 435-451, Oct.-Dec. 2016, doi: 10.1109/TAFFC.2015.2462830