A Hybrid Model to Improve Occluded Facial Expressions Prediction in the Wild during Conversational Head Movements

Arvind K. 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 expression, 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; 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 [1], there is a growing need of social robotics for elderly care and healthcare [2]. To show empathy, interact, and converse with humans, social robots need to understand human emotions and pain [3], [4] in the wild i.e. emotions are derived from the living systems in real-time scenarios, such as attending elderly patients or helping a frustrated elderly person in a home setting [5], [6].

Predicting emotions in the wild is complex and requires multimodal multimedia analysis involving dialogs [7], voice-modulation (including timed silence) [8], gestures (including postures, gaze, conversational head and hand gestures, and haptic gestures) [9], facial expressions [10]-[13], pain, and tears. 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 are yet to be achieved. Compared to emotions exhibited in dialogs, utterances and gestures, facial expressions are exhibited more involuntarily [4], [10].

Multimedia analysis of facial expressions requires a sequence of video frames, dimension reduction, intelligent image analysis, and analysis of the intensity of facial expressions. In cognitive psychology, two approaches are used to study facial expressions: basic six emotions (anger, disgust, fear, happiness, sadness, and surprise) popularized by Ekman and others [14]; Valence-based Plutchik's wheel of emotion that relates positive and negative emotion classes in multiple intensity levels [15]. Computational analysis is currently limited to recognizing six basic emotions [4], [10], [11] due to tractability of the underlying problem and explicit correspondence of basic emotions to facial muscles modeled by facial action units [14].

Previous studies are mostly limited to the frontal facial view [11] or static aligned poses [16] using curated databases [17]-[20] showing nonoccluded pure facial expressions in proper lighting conditions. In recent years, many researchers have suggested techniques to handle information loss caused by partial occlusion due to external face-obstructing objects, such as eye-glasses, hats, scarfs, and medical masks; hand gestures; hair and mustaches; ambient lighting conditions [21]-[29]. These schemes are based on reconstruction of small patches of a partially occluded face using nonoccluded (or global) facial texture. None of these techniques are 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.

In real-life scenarios, the face continuously moves during a conversation [9], [30], [31] based upon 1) conversational gestures, such as argumentation, interrogation and denial [9]; 2) intensity of emotion [15]; 3) multi-party interactions; changing ambient lighting conditions and shadows with head-movements during conversational gestures. Head rotations stochastically occlude feature-points causing information loss hindering accurate facial-expression classification.

Experiments with CNN-based model [13], as described in section 5, show that facial expression prediction drops by 10-

 ISBN: 978-1-61208-882-2

20% for partial occlusion (less than 45° rotation) and by 30-50% beyond 45° rotation.

Recent augmentation of CNN-based modeling with Long Short Term Memory (LSTM) and transfer learning improves temporal context and maps real-time movement to the nearest alignment of static CNN model to improve the prediction [31] during head-movement. However, they do not handle extreme information loss beyond partial occlusion and do not exploit facial symmetry.

CNN-based models need to be augmented with temporal contexts and 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-rotation, only one side of the face is occluded, and facial symmetry can be used to reconstruct the occluded discriminatory feature-points knowing the coordinates of their counterparts on the nonoccluded side.

This research improves facial-expression analysis for face under motion by utilizing facial symmetry [32] along the vertical major axis. Facial symmetry has been used to estimate the coordinates of missing discriminatory feature-points using their nonoccluded counterparts [33], [34]. Prediction is based upon 1) inherent symmetry of the face around the vertical axis of the face; 2) noted differences between the symmetrical points and the actual geometric feature-points from the previous frames.

The proposed hybrid model augments the CNN-based model [13] with a symmetry-based geometric model proposed in this paper. The hybrid model uses CNN-based prediction for the nonoccluded or partially occluded space and the symmetry-based geometric model beyond partially occluded space. The proposed geometric model provides motion continuity and temporal context to the CNN model 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 the partial occlusion.

The overall roadmap of this paper is: Section 2 describes the related work. Section 3 describes background concepts about facial features. Section 4 describes the proposed symmetry-based geometric model. Section 5 describes the implementation and experimental results. Section 6 concludes the paper.

II. RELATED WORK

Related work can be classified as: 1) handling occlusion for improper lighting conditions, hand-gestures and external objects [24]-[32]; 2) analyzing emotions in the wild [6]; 3) a combination of CNN, LSTM and transfer learning to map continuous motion to the corresponding CNN [31].

To handle the occlusion caused by external objects, researchers have used fixed pose alignments [16], hybrid models training on occluded and nonoccluded samples and using nonoccluded features-space as a guidance to predict texture of occluded patches [24], combination of sparse representation and maximum likelihood estimation [25], a combination of Gabor filter and local binary pattern to derive the texture of occluded patch [26], deep structure recognition [27], a combination of feature histogram, dimension reduction and support vector machine [28], Gabor filter and co-occurrence matrices [29], combination of global and local textures with CNN and attention [30], use of LSTM auto-encoders [31], and Bayesian networks [32].

The above schemes combine information from previous images or texture-pattern from nonoccluded space, and dynamic weighting of texture-patterns to reconstruct occluded patches using well curated datasets [17]-[20]. 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 [13]. 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 [23] required for analyzing facial expression during head rotation. 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.

Compared to other schemes, the proposed geometric model significantly exploits facial symmetry [33] - [35] to recover occluded feature-points during extreme head rotations. 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 under extreme head-rotations. In our scheme, the availability of discriminative feature-points supports multimodal analysis of head-gestures and provides explanation capability.

III. BACKGROUND

A face has two types of 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 [10]. Figure 1 illustrates various feature points.

A face has six major *fixed points*: two ends of the left and right eyes; bottom of a nose; middle point between two eye-brows above the nose-tip. There are 14 major active points: 1) three points on each brow; 2) two middle points of lips; 3) two endpoints of the mouth; 4) two middle points in each eye.

 ISBN: 978-1-61208-882-2

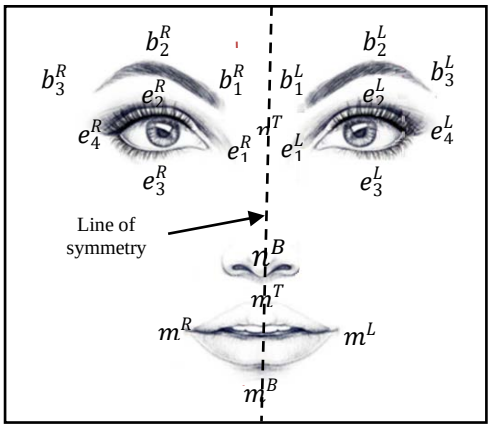

Figure 1. Facial feature points with symmetry

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

A. 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.

B. Facial Symmetry

Facial features have an anatomical symmetry at the muscle level around the vertical axis. This symmetry causes similar changes on both sides of a face for most facial-expressions.

C. Occlusion and Head Movement

In a real-world situation, the head rotations are observed every 5 - 7 degrees [35]. 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 reducing computational overhead. 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.

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.

IV. Proposed 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 t 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). Figure 2 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 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 $n^B n^T$.

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*.

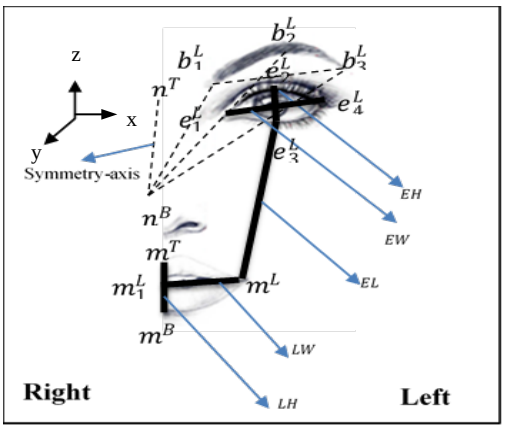

Figure 2. Facial feature-vectors

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 Figure 2). 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.

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

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 $1 \mp \epsilon$.

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). 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 oblique 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*. Overall, these line-segments cover FAUs 1, 2, 4, 5, 6, 7, 8, 9, 10, 12, 14, 15, 16, 17, 20, 23, 26 and 27 involved in six basic facial expressions. The overall correspondence is summarized in Table I.

TABLE I. LINE-SEGMENTS

Line-ratio	Norm. 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

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 II.

TABLE II. 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
$\|\, b_1^L\ b_3^L \,|_x$	4	anger, disgust, fear, sadness
$\|n^B b_1^L\,|_z$	1, 4, 9	anger, disgust, fear, sadness, surprise
$\|n^B b_2^L\,|_z$	4, 5	fear and surprise
$\|n^B b_3^L\,|_z$	2	fear
$n^B n^T$	Used for vertical normalizations	
EW, $\|n^T n^B\|$	Invariant with head-rotation	

C. FAU Correspondence

Table III 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. All the FAUs involved in basic facial expressions are derived using these conditions.

TABLE III. 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	$R_n^{EL} > R_m^{EL}$

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).

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} correspond to the activation of FAU 4 (eye-brow lowerer).

V. IMPLEMENTATION AND EXPERIMENTATION

RaFD database [20] was used for training and comparison of results between geometric modeling and CNN-based model. For the online video capturing, three frames per second were used for the facial expression analysis. Epochs of 200 frames were used because the experimental data show that the accuracy of the facial expression recognition stabilizes around 200 frames.

A. CNN Architecture

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 sub-classification of textures. After each convolution layer, there is a max-pooling layer for the subsampling of images. Each max-pool layer has a 2 × 2 pixel window.

After applying the Locality-Sensitive Hashing (LSH) [12] and Gabor filter [13], the processed images are passed to the network of convolution layers through the input layer. LSH is a dimension reduction technique that maps the pixels with similar values in the same bucket. Gabor filter preserves the texture directionality. The hidden layers extract facial features and reduce the dimensions. The fully connected layer combines the matrix-derived after the last hidden layer into one vector, and the Softmax layer extracts the output from the vector.

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 is 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. Database and Video Processing

RaFD dataset was used for measuring the performance of the CNN-based model for various static alignments in different poses [20], [36]. Compared to other curated facial expression databases u [17]-[19], RaFD gives comprehensive facial-expressions for 67 models (for both genders) with multiple camera angles and adjustment of lighting conditions.

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* as defined in (2).

$$Recall = \frac{true\ positive}{true\ positive\ +\ false\ negative} \quad (2)$$

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

C. Performance Evaluation and Discussion

Table IV illustrates CNN based prediction, even for a cured RaFD database, deteriorates quickly due to the unavailability of discriminatory feature-points on the

 ISBN: 978-1-61208-882-2

occluded part of the face. The deterioration varies from 48% for sadness to 41% for happiness for complete occlusion.

TABLE IV. RECALL IN CNN MODEL WITH RADB DATASET

	Right complete occlusion	Right part occl.	Front no occl.	Left part occl.	Left complete occlusion
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 V. CONFUSION MATRIX - CNN MODEL (FRONTAL POSE) IN WILD

	sad. %	disg. %	ang. %	fear %	sur. %	happ. %	neu.
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
happines	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 VI. RECALL IN HYBRID MODEL IN WILD

	Right complete occlusion	Right part occl.	Front no occl.	Left part occl.	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%
happines	75%	79%	87%	81%	77%

Comparison of Table IV and Table V 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. 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 [30].

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 IV and Table VI 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.

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.

VI. CONCLUSION AND FUTURE WORK

Head-motions during conversational gestures and multi-agent interactions cause extreme occlusion of 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.

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 [30].

REFERENCES

[1] 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.

[2] 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, Daegu, South Korea, 2019, pp. 669-670.

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

[4] 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.

[5] F. Rothganger, “Computation in the Wild,” https://www.osti.gov/servlets/purl/1644432, [accessed date; June 12, 2021.

[6] 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, 2016, pp. 163-172. doi:10.1007/s12193-015-0210-7.

 ISBN: 978-1-61208-882-2

[7] 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.

[8] 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, Singapore, Sept. 2014, pp. 223-227.

[9] A. Singh and A. K. Bansal, “Towards Synchronous Model of Nonemotional Conversational Gesture Generation in Humanoids, Computing Conference, London, UK, July 2021, in press.

[10] 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, 1991.

[11] 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.

[12] 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, Vienna, Austria, Aug. 2014, pp. 211-219.

[13] M. Ghayoumi and A. K. Bansal, “Emotions in Robot using Convolutional Neural Network,” International Conference on Social Robotics (ICSR 2016), Kansas City, KS, USA, Nov. 2016, pp. 285-295.

[14] 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.

[15] R. Plutchik, “Emotion: A Psychoevolutionary Synthesis,” New York, NY: Harper & Row, 1980.

[16] 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, Oct. 2016, pp. 2110-2122, doi: 10.1109/TPAMI.2015.2505301.

[17] 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), San Francisco, CA USA, June 2010, pp. 94-101, doi: 10.1109/CVPRW.2010.5543262.

[18] 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.

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

[20] 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, Dec. 2010, doi: 10.1080/ 026999 30903485076.

[21] 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, Apr. 2018, doi: 10.1145/3158369.

[22] 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), Beijing, China, July 2014, pp. 1285-1290.

[23] 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, Nanjing, China, July 2014, pp. 4064-4068.

[24] 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.

[25] 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, 2016, pp. 1298-1302.

[26] 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, Jabalpur, India, Dec. 2015, pp. 347-351.

[27] 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.

[28] F. Zhao, J. Feng, J. Zhao, W. Yang, and S. Yan, “Robust LSTM-Autoencoders for Face De-Occlusion in the Wild,” vol. 27, no. 2, pp. 778-790, Feb. 2018.

[29] Y. Miyakoshi and S. Kato, “Facial emotion detection considering partial occlusion of face using Bayesian network,” IEEE Symposium on Computers & Informatics, Kuala Lumpur, Malaysia, March 2011, pp. 96-101.

[30] 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.

[31] 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.

[32] 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.

[33] 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), Portland, OR, USA, June 2013, pp. 217-222, doi: 10.1109/CVPRW. 2013.39.

[34] 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.

[35] 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.

[36] 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, doi:10.3390/ijerph17207420