Towards Formal Multimodal Analysis of Emotions for Affective Computing

Mehdi Ghayoumi
Artificial Intelligence Lab
Department of Computer Science
Kent State University
OH, USA
mghayoum@kent.edu

Maha Thafa
Artificial Intelligence Lab
Department of Computer Science
Kent State University
OH, USA
mthafar@kent.edu

Arvind K. Bansal
Artificial Intelligence Lab
Department of Computer Science
Kent State University
OH, USA
akbansal@kent.edu

***Abstract*—Social robotics is related to the robotic systems and human interaction. Social robots have applications in elderly care, health care, home care, customer service and reception in industrial settings. Human-Robot Interaction (HRI) requires better understanding of human emotion. There are few multimodal fusion systems that integrate limited amount of facial expression, speech and gesture analysis. In this paper, we describe the implementation of a semantic algebra based formal model that integrates six basic facial expressions, speech phrases and gesture trajectories. The system is capable of real-time interaction. We used the decision level fusion approach for integration and the prototype system has been implemented using Matlab.**

Keywords- Affective computing, Emotion recognition, Human-machine interaction, Multimedia, Multimodal, Decision level fusion, Social robotics.

I. INTRODUCTION

In the Human Compute Interaction (HCI) researches and studies, facial expression, speech and body movements have the major roles [8, 18 and 24]. Due to the aging society and increasing cost of health care, elderly care and assisted living, there has been significant interest in the development of social robotic systems that can interact with humans through the use of sensors. The social-robotic system can be humanoid, computers or intelligent machines as in Internet of Things who will interact with humans in the daily life. Interacting with humans requires understanding emotions [6] and emotions are based on a person's state of mind and partially regulated by personality, context and conditioning. Emotion is a language for communicating by feelings and it includes approval and disapproval to robotic systems. Interactive emotions [8, 16, and 18] are a subclass of human emotion analysis that humans use to interact with each other in close proximity. There are many interactive emotions that a person can express to machine during interaction, such as happiness, anger, embarrassment, surprise, rage, disappointment, confusion, elation, depression, approval and disapproval. Interactive emotions are expressed using a combination of verbal and nonverbal modes such as facial expressions [7, 10,], speech [17] including silence, gesture including body-posture and body-motion. Single mode may not give the emotion completely, or may be unavailable during emotive interaction. For example, the face may be occluded by the hands during sadness when a person is in deep pain or is crying. A person in shock or deep pain may not utter a single word. In the early stages of social robotics, most of the human-computer interaction in the service industry will involve brief commands or query by the human, and the robots will play a subordinate role rather than as companion role. Most of the emotion recognition will be limited to the integration of:

1) Facial expressions,
2) The limited amount of speech commands and emotional phrases to provide as an approval, disapproval, encouragement of a robot response or action, and
3) Gesture analysis of the upper body part involving head, hand, fingers, eyes, and lips.

The speech commands may be restricted to commands like "yes", "no", "don't like", "very good", "I am happy" etc. Some of these commands may have limited speech attributes such loudness showing disapproval or anger. Speech analysis can be done by a combination of text-to-speech conversion to understand the emotional phrases, and fast Fourier transform can be used to derive the variation of speech features such as energy, amplitude envelope, pitch during emotional variation. Real-time facial expression analysis in a video analysis where emotions and emotion transition can be studied by frame-to-frame analysis of facial expressions. Gesture analysis is done by analyzing video analysis and depth analysis as in Kinect based systems.

Currently, computational systems are limited to analyzing a single mode of emotion expression such as facial expression, speech, and (to some limited extent) multimodal analysis [22].

DOI reference number: 10.18293/DMS2016-030

The current integration of multimodal analysis systems of interactive emotions, lack:

1) A formal model to combine multiple modes such as facial expressions, speech analysis and gestures,
2) Complete catalog of upper body gestures, and
3) Capability of real-time analysis of facial-expressions and gestures.

In biometric systems, multimodal systems have some advantages which make the system accuracy and performance higher. Here we are using this model to achieve better results [23]. This research effort is in the direction of real-time integration of multimodal analysis system to derive emotion during HRI. A fusion module combines the weighted scores derived from each mode to derive the best emotion. The major contribution, here are:

1) Implementation of a real-time facial expression system based upon integration of geometric features, facial action units and facial symmetry [27],
2) Gesture recognition systems using fuzzy values,
3) Emotional Phrase lookup module,
4) Weighted score based Integration of a multimodal system based upon an abstract model of multimodal emotion analysis.

The rest of the paper is organized as follows. Section 2 describes background. Section 3 describes the overall architecture. Section 4 explains the speech recognition system and facial expression analysis will be explained in the section 5. Section 6 demonstrates the gesture modeling. Section 6 illustrates the implementation and performance results. Section 7 demonstrates the related works and the last section concludes the work, and describes the future directions.

II. BACKGROUND

This section describes the background material related to facial expression, speech analysis and gesture recognition and describe basic mathematical concepts needed for abstract modeling of the emotions.

A. Components of Emotion Recognition

There are three popular psychological theories of emotions: *James-Lange theory* [19], *Cannon-bard theory* [20] and *Schater-singer* theory [21]. *James-Lange* theory states that the mental state in response to the reactions which caused by external stimuli is emotion. *Cannon-Bard* theory is based upon anticipation rather than as a reaction to specific action. *Schachter-Singer* theory states that encountering an emotion requires both an interpretation of the bodily response as well as specific circumstance at a specific moment.

Also, there are three major classes of emotions:

1) Basic emotions,

2) Emotions that having same basic class, but having different intensity,

3) Mixed emotions that are a combination of one or more basic and/or mixed emotions.

Although, there are some disagreements among researchers, and a popular computational theory of Ekman [22] identifies six basic emotions: *happiness, sadness, surprise, disgust, anger* and *fear*. An example set of emotions having same basic class, but different intensities is {relaxed, happy, delighted, and euphoric}. Another set is {upset, anger, rage} etc. An example of mixed emotion is {amazed} that is a combination of {surprise and happiness} or {envy} which is the combination of {sadness and anger} or {despair} which is the combination of {fear and sadness}. In general, Facial Expressions have been done using these types of systems:

1) Facial Action Coding System (FACS) based on the simulation of facial muscle movement,
2) Geometric Features Modeling (GFM) based upon the movement of major feature-points of the face such as dynamic change in location endpoints and curvature of the mouth, eye, lips, forehead furrows and space between eyebrows.

The unit of FACS is an Action Unit (AU) that involves a segment of a muscle in facial expression. There are 17 major AUs involved in basic facial expressions. Examples of AUs involved in facial expressions are: inner brow raiser, outer brow raiser, brow lowered and drawn together, upper eye-lid raised, cheek raised, upper lip raised, lip corners pulled down, etc. The major geometric feature points, involved in facial expression analysis are given in Figure 1 which these features-points include:

1) 3 eyebrow points in each of the eyebrows:
$$b_1^L,\ b_2^L,\ b_3^L,\ b_1^R,\ b_2^R,\ b_3^R$$
2) 2 endpoints of eyes in each of the eyes:
$$e_1^L,\ e_2^L,\ e_1^R,\ e_2^R$$
3) Middle points eye-lid in each of the eyes:
$$el_L \text{ and } el_R$$
4) 2 endpoints of nose:
$$n^T \text{ and } n^B$$
5) 2 endpoints of mouth:
$$m^L \text{ and } m^R$$
6) 2 middle points of the mouth based on top and bottom lips:
$$m^T \text{ and } m^B$$
7) Chin-point denoted as:
$$ch\ .$$

The points shaded in dark black- $e_1^L, e_2^L, e_1^R, e_2^R, n^T$ and n^B do not move, and act as reference-points. Remaining spotted-points move with emotions, and their displacement is used to derive the facial expression.

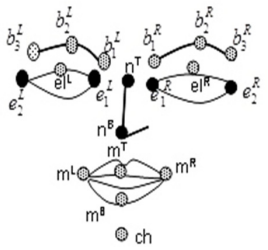

Figure 1. Major feature-points on the face

Emotional speech has multiple features such as phonemes, emotional phrases, amplitude, syllable envelope, pitch, rhythm, quantile and silence. Phonemes are the basic units of speech. During emotional interaction, pitch, amplitude, syllable envelope, duration of silence and utterances change significantly, act as parameters for the recognition of interactive emotions. Gesture is a nonverbal communication using perceptible bodily actions such as body-postures and body-part movements, including movements of the head, torso, hands, face and eyes. Different components of the emotions are measured using different sensors. Facial-Expression uses image analysis techniques to identify the movement of facial feature points, speech analysis uses wavelet analysis, FFT analysis, morphology analysis, text-to-speech conversion for phoneme detection and dictionary lookup to identify phrases. Gesture recognition requires image analysis to derive postures and video-frame analysis to derive motion of various body parts such as head, arm, eyes, hand, palm, fingers. The posture and motion are modeled as fuzzy values to reduce the computational space. The motion of the body parts can also be derived using skeletal and depth analysis used in Kinect.

B. Mathematical concepts

The Fuzzy values map a large value-space to a smaller finite space. The major advantages of the use of fuzzy values are:

1) Reduction of the computational complexity
2) Nearness to human perception and
3) Tolerance from the sensor noise.

We use two types of fuzzy sets:

1) Discrete fuzzy set, and
2) Ordered fuzzy sets.

A discrete fuzzy set has values that have no relationship that shows transitivity. For example, a head posture can be {rotated-left, rotated-right, normal, tilted-left, tilted-right, looking-down, looking-up}. An ordered fuzzy set shows transitive relationship between the values, and is used to model motion intensity in gesture analysis for better classification of emotion. For example, the speed of a head-motion can be modeled as {still, slow, normal, fast, very fast}. The values in the fuzzy set can be mapped onto the ordinals 0… 4:

Still →0, Slow→1, Normal→2, Fast→3 and Very Fast→ 4 (1)

The use of this mapping allows the use of comparison operators on ordered fuzzy sets. Cartesian product of the N sets returns a set of N-tuples such i^{th}-field of an element is a member of the i^{th} set as shown:

$$X_1 \times \ldots \times X_n = \{(x_1, \ldots, x_n) \mid x_i \in X_i \forall\, i = 1, \ldots, n\} \quad (2)$$

Two domains can be joined using:

1) Product-domain that uses the Cartesian product $A \times B$, or

2) A sum - domain that uses disjoint-union $A + B$, or

3) *Function Domain* mapping on lifted domains f: A⊥ B. Where ⊥ is the bottom symbol used to catch all ill-defined mappings.

III. OVERALL ARCHITECTURE

Overall architecture (see Figure 2) has six major modules:

Unit 1 - Facial Expressions Subunit (FE)

The subunit takes a video-clip that is a sequence of frames, and extends the integration of FACS + geometric feature analysis technique for real-time basic face-expression recognition to derive the ranked subset of facial expressions for each frame in the video-clip. The analysis of a frame may result in more than one facial expression due to the:

1) Partial or full occlusion of the face due to gesture or head rotation,
2) Transition of emotions,
3) Inherent accuracies in the facial expression technique,
4) A variation of the facial expression due to the situation, personality or culture,
5) Low emotional intensity.

The outcome of this facial-expression analysis gives a sequence of subsets of derived possible emotions with the matching score of the form:

$$<FE_1,\ldots, FE_j, FE_{j+1},\ldots, FE_N> \quad (3)$$

Where N is the number of frames in the clip, FE_i is a subset of rank facial expressions of the form

$$\{(e_1, s_1)\ldots (e_m, s_k)\}\ \text{For}\ k \geq 1, e_i \in \Sigma, s_i > \text{threshold and}\ s_i > s_{(i+1)} \quad (4)$$

Which e and s are the emotion and its corresponding score respectively.

Unit 2 Gesture Fuzzy Parameterization Module (GFP)

The model measures different postures and motion intensity and frequency of different emotional gesture patterns.

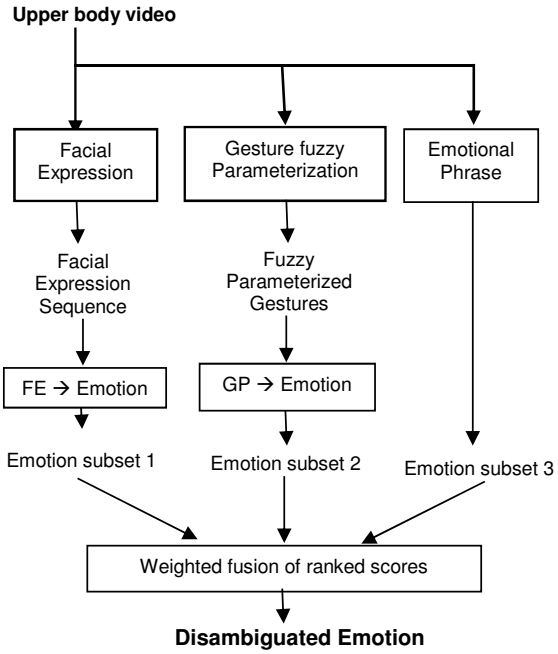

Figure 2. Overall architecture of multimodal fusion

The different postures are: body-posture, head-posture, shoulder-posture, hand-postures, palm-posture, finger-postures and eye-gaze and the various-motions are: head-motion, arm-motions, eye-motions, finger-motions. The details of the fuzzy parameterization and emotional head-motion gesture is given in Section 5.

Unit 3 - Emotional Phrases Module

Emotional phrase modules use a hash function to generate the index, and stores multiple emotions with a fuzzy intensity value. Once a phrase is recorded, then text-to-speech converter is used to derive the text. Individual words are looked up in the user-specified dictionary to remove the noise in text-to-speech conversion. The speech analysis system is used to derive the relative energy level. The energy level is parameterized to a fuzzy value, and the hash function is used in the derived text to identify the index of the speech. Using this index, the corresponding set of emotions that closely matches the intensity levels are derived.

Unit 4- FE (Emotion Module)

Many adjacent frames will have the same subset of facial expressions until the facial expression changes. We call this frame as the emotion transition point. Thus:

$$<FE_1, \ldots, FE_j, FE_{j+1}, \ldots, FE_N> \rightarrow <(E_1, d_1), (E_2, d_2)\ldots, (E_M, d_M)> \quad (5)$$

Where $E_I (1 < I < M)$ is a subset of Σ, and d_I is of the form start-frame: end-frame. The term d_I is used:

1) To reconcile the emotion in the fusion-module, and
2) To derive the duration of emotion to match with the duration of emotion derived from gesture analysis and emotional phrase analysis.

Unit 5- GFP (Emotion Module)

This module takes the fuzzy parameterized values of different body parts and their motion, concatenates them into a long string, and creates a string, and performs a similarity-based search in K-dimensional space where K is the number of fuzzy-valued component to identify a possible set of emotions. The attributes of fuzzy-vector representation of head-motion trajectory is hashed to derive the possible set of emotions based upon gesture. The duration of the body-motion is noted like unit 4, and its output is also on the form:

$$<(E_1, d_1), (E_2, d_2)\ldots, (E_M, d_M)> \quad (6)$$

Where $E_I (1 < I < M)$ is a subset of Σ, and d_I is of the form start-frame: end-frame.

Unit 6 - Weighted Fusion of Emotion Module

The role of the weighted fusion module is to:

1) Fuse the information of the ranked emotions from the three modules to reduce ambiguity and improve ranking scores,
2) Derive the duration of emotion.

The input of the FE (emotion module (unit 4), and GP (emotion module (unit 5) and unit 3 are a sequence of set of rank emotion with the duration. Under the assumption, that emotions are expressed involuntarily in the facial expressions first, the start frame of the facial expression should occur before followed by the emotions derived from other two modules. To handle the issues that emotions may not be expressed by one or more modules, the weight is dependent upon:

1) Availability of the emotions from the specific mode,
2) The noise level.

For example, initial weight is w_1, w_2 and w_3 for the fusion of the corresponding modes, the weights $w_k (1 \leq i \leq 3)$ is altered by:

$$\left(\sum_{i=1}^{i=3} w_i \Big/ \sum_{i \neq j, i=1}^{i=3} w_i\right) \quad (7)$$

Since one of the modes is missing. The fusion is performed by:

1) Multiplying the ranked score of each emotion by the corresponding weight,
2) Adding the scores of the same derived emotions from different weights, and
3) Sorting the emotions by the descending order of the scores, and picking up the emotion with the highest score.

If the top two scores are very close, then it can be a case of emotional transition or mixed emotion.

IV. Facial Expression Analysis

We extend the integration of FACS system interaction and geometric feature analysis [8] to make the facial expression analysis by using facial symmetry and invariance under head-motion. There are 13 moving-points (11 active points and 2 passive points) and 6 references-points. FACS system analysis has been used to derive the features-points that are significant during the expression of a specific facial expression. For example, for a surprise the all eyebrow points are uniformly raised; for happiness mouth corners are stretched, the eye-lid point gets lowered; for anger distance between eyebrows becomes smaller, inner eyebrow points get lowered. These FAUs have been translated to the corresponding feature-point movements as given in Table 1. We denote vertical-up motion by ↑, vertical-down motion by ↓, horizontally stretched outwards by '↔', horizontally compressed inwards by '↔', oblique-stretched downwards by '↘', oblique-stretched upwards by '↗'. If the emotion is symmetric, then the subscripts L and R have been omitted. If the movement is optional or shows higher intensity increase then it has been placed within the square brackets. Conjunction has been shown using concatenation. Essential feature-point have been within parenthesis () separated by ','. At least one of the essential feature point motion has to be present for the emotion to occur. Scores are associated with the presence of each feature-point motion.

TABLE1. Feature Point displacements

Facial Expressions	Major Feature-points displacements
Anger	$(e_1\leftarrow el\uparrow) + [e_2\uparrow] + [m^T\uparrow m^B\uparrow]$
Disgusted	$(m^T\uparrow ch\uparrow) + [\{m^L, m^R\}\downarrow] + [m^B\uparrow]$
Fear	$(e_1\uparrow, m^L\downarrow m^R\downarrow) + [m^T\downarrow] + [e1\leftarrow]$
Happiness	$(m^L\nearrow m^R\nearrow, M^T\uparrow m^B\downarrow ch\downarrow m^L\leftrightarrow m^R\leftrightarrow)$
Sadness	$(el\downarrow m^L\leftrightarrow m^R\leftrightarrow) + [ch\downarrow]$
Surprised	$(e^1\uparrow e^2\uparrow e^3\uparrow el\uparrow ch\downarrow) + [m^T\uparrow m^B\downarrow]$

For each feature point, we measure the displacement distance and the direction of the displacement. Thus the derivable facial expressions are mapped to a vector of (displacement-distance ratio, direction).

Direction is a discrete-fuzzy set with six possible values:

1) Vertical-up,
2) Vertical-down,
3) Horizontal-compressed-inwards,
4) Horizontal-stretched-outwards,
5) Oblique-stretched-upwards, and
6) Oblique-stretched-downwards.

Facial expressions can be occluded due to:

1) Gestures such as hand covering face in case of sadness,
2) Lighting conditions, and
3) Head rotation or tilt.

In order to complete the information, we use the facial symmetry around nose to fill in the information about those facial expressions that show symmetry such as happiness, anger, surprise, sadness and fear. Disgust may shows asymmetric features. In order to variation of the displacement projection due to head motion, we use the distance-ratio (point-displacement from the relaxed state / distance between reference-points) which go with similar transformation.

For example, to keep the horizontal displacement we use distance between the two outer-eye corners: e_2^L and e_2^R in the denominator of the ratio; and for the two noses displacement. We use the distance between the two noses-points and n^B.

There are two types of motion-points:

A. Points that move in only in vertical up-down direction, such as:

$$e_1^L, e_2^L, e_1^R, e_2^R \text{ and}$$

B. Points that move in all four directions: vertical up-down and horizontal inside-outside motion such as:

$$el^L, el^R, m^L, m^R, m^T, m^B, ch$$

For the points that show motion in vertical up-down direction have only one entry in the *facial-expression ratio vector*, and points that show motion in up-down and stretch-compression mode have two entries in the vector. Based upon the emotion, some of the entries may not change during that emotion.

For example, in surprise, only, $el^R, e_2^L, e_3^L, e_2^R, e_3^R, m^T, m^B$ and ch change. This characteristic of facial-expression ratio-vector provides invariance against head-motion as well as specific characterization of facial-expressions.

V. Speech Analysis

Embedding the component of emotion processing into existing speech systems makes them more natural and effective. Several approaches to recognize emotions from speech have been reported. In a conversation, non-verbal communication carries an important information like the intention of the speaker. In addition to the message conveyed through text, the manner in which the words are spoken, conveys essential non-linguistic information. The same textual message would be conveyed with different semantics by incorporating appropriate emotions. Spoken text may have several interpretations, depending on how it is said. For example, the word 'OKAY' in English, is used to express respect, disbelief, agreement, and disinterest. Therefore, understanding the text alone is not sufficient to interpret the semantics of a spoken utterance. However, it is important that, speech systems should be able to process the non-linguistic information such as emotions, along with the message. Choosing suitable features for developing any of the speech systems is a crucial decision.

We have three important speech features, namely:

1) Excitation source,
2) Vocal tract system, and
3) Prosodic features.

VI. Gesture Analysis

Gesture parameterization has two modules:

a) Deriving the posture of upper-body parts and their motions,

b) Mapping fuzzy to actual values to reduce the search space.

These fuzzy values are concatenated so that all the values from discrete sets are concatenated together, and all the values from the ordered sets are grouped together. This separation is necessary because mismatch in discrete sets leads to failure, while mismatch in ordered sets is permissible. Abstract modeling of emotion requires functional mapping of Cartesian product of different components to derivable emotions. Since all the component tuples may not map to valid emotional elements in the emotional domain, we make the emotion domain a lifted domain by introducing a bottom symbol $\perp$ in the set of well-defined emotions. The lifted domain allows for catching the error conditions when the tuple of fuzzy component values do not map to any specific emotion.

Fuzzy values are calculated using statistically derived thresholds. There are two types of sets:

1) Discrete sets where the values are not ordered and ordered sets,
2) Ordered sets are used in modeling the extent of posture variation and intensity of the motions of various gestures.

The parameters of the motion are:

a) Start-position,
b) End-position,
c) Frequency,
d) Speed,
e) Attack,
f) Relaxation.

The attack is the rate of change of speed until the motion attains the peak speed, relaxation is the rate of reduction of speed to the speed reduces from the peak speed to no motion. The mapping of the components is described by equation (7).

$$\text{Posture}_1 \times \ldots \times \text{Posture}_M \times \text{Motion}_1 \times \ldots \times \text{Motion}_M \rightarrow \text{set of possible emotions} \quad (7)$$

VII. Implementation and Performance Results

The implementation can be divided into 3 major steps as follows:

A. *Implementing Facial Expression Sequence Analysis*

We extract the main parts of the face such as eyes, eyebrow, nose and mouth, then find the key points in each segment as shown in the figure 1 then we extract the feature vectors from extracted key points and train the networks.

B. *Implementing Head Gesture Analysis*

Human head movement is very important in general conversation and communication. Despite the influential role of the head gestures, very little research has examined the gesture's role in the robot-human interaction process. In software module, the pose of the human head is estimated with a constraint that the human head is a 3 DOF rigid object which has yaw, pitch and roll movements. We have used geometric head pose estimation algorithm which estimates the head pose through a standard webcam of the computer. The details of this algorithm can be found in [12].

C. *Implementing Emotional Phrase Matching*

A cepstrum is obtained by computing the Fourier Transform of the logarithm of the spectrum of a signal. There are different kinds of cepstrum such as complex cepstrum, real cepstrum, phase cepstrum and power cepstrum. The power cepstrum is used in speech synthesis applications and here we use it. The approach based on decision-level fusion obtained. The performance of the classifier was 94.6%, both for the best probability and for the majority vote plus best probability approaches.

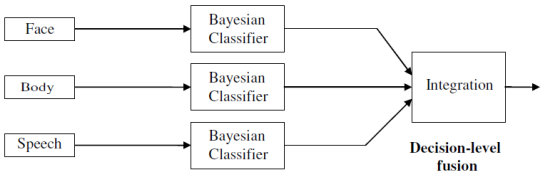

Figure 3. The decision level fusion

Table 2 shows the performance of the system with decision level integration using the best probability approach. Anger has the emotion recognized with highest accuracy.

TABLE 2. Decision level integration with the best probability approach

Anger	Happy	Sad	Surprise	Disgust	Fear	
98.3	0	0	0	4.3	0	**Anger**
0	**95.4**	0	7.2	0	0	**Happy**
3.1	0	**92.1**	0	2.7	2.2	**Sad**
9.3	10.4	0	**87.5**	2.3	5.8	**Surprise**
7.2	0	4.1	3.3	**90.2**	4.2	**Disgust**
0	0	0	12.2	11.1	**83.3**	**Fear**

VIII. Related Works

There are many related works in the facial expression analysis [1, 10, 25 and 26], gesture analysis [2, 4, 12, and 18] emotion recognition in speech [9] and multimodal fusion [3, 4]. Castellano et al. [7] extended their work to multimodal framework integrating face-expression, body-gestures, and speech. A Bayesian classifier was used for feature level fusion and decision level fusion. A comparison between unimodal, bimodal, and multimodal classification showed that multimodal classification is better. There is an additional need to identify features that are relevant to the dynamics of expressive emotions, however, their study is limited to identifying eight emotions [7]. We have been influenced by the research to derive from the research to analyze facial expressions based upon action units and map action units based movement to geometric feature-point movements [8]. The use of geometric feature-points and fusion allows for better accuracy in our research. In addition, we use abstract model for gesture analysis. Our geometric point movement based upon study of facial action units is generally enough to analyze finer classification of emotions and mixed emotions. Many interesting works about audio-visual fusion/mapping has been proposed for multimodal information processing. For instance, speech based facial animation [13], and audio-visual based emotion recognition [14]. There is also some work for head motions [15] and body gestures [16], however, most of them just focused on the gesture recognition.

IX. Conclusion and Future Works

In this paper, we have described a detailed methodology and an initial prototype implementation of real-time multimodal fusion to derive interactive emotion for interaction with social-robots and intelligent machines with limited emotional phrase based interaction. The proposed integrated system has many novelties such as: an abstract model of fusion based upon a semantic algebra that the maps Cartesian product of different components to derivable emotions, the use of invariant displacement of geometric feature-points to identify facial-expressions, and Gestures based upon head-trajectory and fuzzy values of other upper body parts to reduce the search space. Currently, the gesture based system is limited to, image analysis of feature-points in the head and hand to derive posture. We are looking into Kinect based analysis to integrate skeleton based body posture, motion and depth analysis [15] for better accuracy.

References

[1] R. Adolphs. "Recognizing emotion from facial expressions: psychological and neurological mechanisms," *Behav. Cogn. Neurosci. Rev.*, Vol. 1, 2002, pp. 21-62.

[2] V. Bevilacqua, D. Barone, F. Cipriani, G. D'Onghia et al., "A new tool for gestural action recognition to support decisions in emotional framework", *Proceedings of the IEEE Symposium of Innovations in Intelligent Systems and Applications* (INISTA), 2014, pp. 184-191.

[3] G. Caridakis, G. Castellano, L. Kessous, A. Raouzaiou, L. Malatesta, et al. "Multimodal emotion recognition from expressive faces, body gestures and speech"; *Artificial Intelligence and Innovations: From Theory to Applications*, Springer Berlin Heidelberg, 2007, pp. 375-388.

[4] G. Castellano, S. D. Villalba and A. Camurri. Recognizing human emotions from body movement and gesture dynamics"; *Affective Computing and Intelligent Interaction*, Springer Berlin Heidelberg, pp. 71-82.

[5] P. Ekman and W. Friesen, "Facial Action Coding System: A Technique for the Measurement of Facial Movement," *Consulting Psychologists Press*, Palo Alto, 1978.

[6] J. M. Fellous and M.A.Arbib,"Who needs emotions? The brain meets the robots", Oxford press, 2005.

[7] L. Gang, L. Xiao-hua, Z. Ji-Liu and G. Xiao-gang, "Geometric feature based facial expression recognition using multiclass support vector machines," *IEEE International Conference on Granular Computing (GRC '09)*, 2009, pp. 318-321.

[8] M. Ghayoumi and A. K. Bansal, "Unifying Geometric Features and Facial Action Units for Improved Performance of Facial Expression Analysis," Proceedings *of the International Conference on Circuits, Systems, Signal Processing, Communications and Computers* (CSSCC 15), pp. 259-266.

[9] C. H. Lim, E. Vats and C. S. Chan, "Fuzzy human motion analysis: A review," *Pattern Recognition,* Vol. 48, 2015, pp. 1773-1796.

[10] S. Mitra and T. Acharya. "Gesture recognition: A survey," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, Vol. 37, 2007, pp. 311-324.

[11] C. M. Lee, S. S. Narayanan, "Towards Detecting Emotions in Spoken Dialog," *IEEE Trans. On Speech and Audio Processing*, Vol. 13, No. 2, 2005, pp. 293-303.

[12]] M. Khan, S. Rehman, Z. Lu and H. Li, "Head Orientation Modeling: Geometric Head Pose Estimation using Monocular Camera," in Proceedings of the 1st IEEE/IIAE International Conference on Intelligent Systems and Image Processing, 2013.

[13] Alberto B, Piero C, Giuseppe RL, Giulio P Lucia" a new WebGL-based talking head," 15th Conference of the International Speech Communication Association, Singapore 2014.

[14] Cowie, R, Douglas-Cowie, E "Emotion recognition in human-computer interaction." IEEE Signal Processing Magazine. pp. 33-80, 2001.

[15] Bo X, Georgiou Panayiotis G, Brian Baucom, Shrikanth S Narayanan, "power-spectral analysis of head motion signal for behavioral modeling in human interaction," I.E. International Conference on Acoustics, Speech, and Signal Processing, 2014.

[16] Welbergen HV, Reidsma D, Ruttkay ZM, Zwiers EJ D, "A BML realizer for continuous, multimodal interaction with a virtual human." Multimodal User Interf 3D4:271-284, 2010.

[17] Oh-Wook Kwon, Kwokleung Chan, Jiucang Hao, Te-Won Lee, "Emotion Recognition by Speech Signals" - INTERSPEECH, 2003.

[18] M. Ghayoumi, A. Bansal, "An Integrated Approach for Efficient Analysis of Facial Expressions", SIGMAP 2014.

[19] A, Walter. "The James-Lange Theory of Emotions: A Critical Examination and an Alternative Theory." The American Journal of Psychology 39: 106-124.

[20] Friedman, B.H. "Feelings and the body: The Jamesian perspective on autonomic specificity of emotion." Biological Psychology 84: 383-393 (2010).

[21] Cotton, J. L. "A review of research on Schachter's theory of emotion and the misattribution of Arousal." European Journal of Social Psychology 11: 365-397 (1981).

[22] P. Ekman, Facial expression and emotion. American, 8 (4): 384-392, 199.

[23] M. Ghayoumi, "A Review of Multimodal Biometric Systems Fusion Methods and Its Applications.", ICIS, USA, 2015.

[24] H. Abrishami Moghaddam and M. Ghayoumi, "Facial Image Feature Extraction Using Support Vector Machines.", Proc. VISAPP, Setubal, Portugal, 2006.

[25] Ghayoumi, M., Bansal, A. K."Multimodal Architecture for Emotion in Robots Using Deep Learning." Future Technologies Conference, San Francisco, United States, FTC 2016.

[26] M. Ghayoumi, A. K. Bansal, "Emotion in Robots Using Convolutional Neural Networks.", ICSR 2016.

[27] M. Ghayoumi, A. K. Bansal, "Real Emotion Recognition Algorithm by Detecting Symmetry Patterns with Dihedral Group.", MCSI 2016.