CIT. Journal of Computing and Information Technology, Vol. 28, No. 2, June 2020, 125–146
doi: 10.20532/cit.2020.1004841

A Survey on Emotion Recognition for Human Robot Interaction

Suhaila Najim Mohammed[1, 2] and Alia Karim Abdul Hassan[1]

[1]Department of Computer Science, University of Technology, Baghdad, Iraq
[2]Department of Computer Science, College of Science, University of Baghdad, Baghdad, Iraq

With the recent developments of technology and the advances in artificial intelligent and machine learning techniques, it becomes possible for the robot to acquire and show the emotions as a part of Human-Robot Interaction (HRI). An emotional robot can recognize the emotional states of humans so that it will be able to interact more naturally with its human counterpart in different environments. In this article, a survey on emotion recognition for HRI systems has been presented. The survey aims to achieve two objectives. Firstly, it aims to discuss the main challenges that face researchers when building emotional HRI systems. Secondly, it seeks to identify sensing channels that can be used to detect emotions and provides a literature review about recent researches published within each channel, along with the used methodologies and achieved results. Finally, some of the existing emotion recognition issues and recommendations for future works have been outlined.

ACM CCS (2012) Classification: Human-centered computing → Human computer interaction (HCI) → HCI theory, concepts and models
Human-centered computing → Human computer interaction (HCI) → Interaction techniques

Keywords: affective computing, social robotics, emotional HRI challenges

1. Introduction

Human-Robot Interaction (HRI) and Social Robotics (SR) are sub-fields of Human-Computer Interaction (HCI) that strive on design, modeling, implementing and evaluating systems for robotics in both controlled settings and real world environments [1]. The main purpose of HRI is development of social robotics systems with capabilities in acting and reacting physically, socially, safely and emotionally with humans. Social robotics extends HRI concept in such a way that the robot could modify its behavior according to the emotion that users express, or it can express a given emotion, resulting in a raise of collaboration with users in the social group. A general framework for HRI system with emotion recognition system is shown in Figure 1 [2]. Different robotic systems that consolidate emotion recognition framework have been built; among these systems is the robot system of Kismet robot. Kismet was made at the Massachusetts Institute of Technology (MIT) by Cynthia Breazeal [3] and it was developed to study how emotion-based robotic system can enhance the performance of interaction with humans.

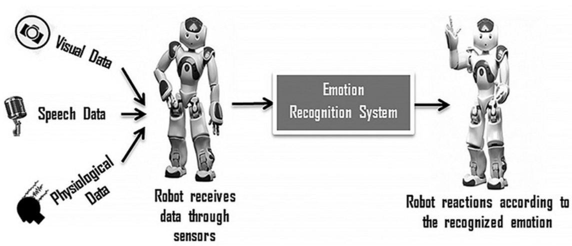


Figure 1. General framework of HRI with emotion recognition.

Enabling robots to understand and shape emotions will evolve their efficiency across a wide number of HRI applications [2]. For example, designing a game playing robot that can recognize the user's emotions of sadness, anger, boredom or joy can increase the success level or fun in the game in a way that the robot can encourage the winner and commiserate the loser. Dang *et al.* (2015) [4] presented a NAO robot system capable of playing "Operation Board Game". The system measures the stress level of the player through collecting player's heart rate during the game. The robot then generates normal or stress alarms based on the detected emotion. Another example of emotional HRI applications is in therapy. For instance, children with autism feel more satisfied around robots than in the presence of humans. Thus, building humanoid robot system that can sense and express emotions will help in improving the capabilities of children with autism to recognize the feelings of others [5]. Salvador *et al.* (2015) [6] programmed "Zeno" robot to show 13 different emotions. Children with autism were asked to guess what emotion the robot displayed. Similarly, Pour *et al.* (2018) [5] proposed a facial based emotion recognition system to be used in autism therapy in which a robot called "Mina" was used for experimental purpose. The authors confirmed the potential of using a robot with facial expressions to improve social skills of children with autism.

However, building robotic systems with emotion recognition ability is a challenging task and requires adaptation of pattern recognition, machine learning and artificial intelligent techniques [3]. Emotional robots must have the flexibility of interacting dynamically in changing environments. Dynamic and strong features of emotions must be extracted to build new models and also models inspired from neurobiology that could adapt to the unsteady environments [7].

One important question that may be asked when building emotion recognition systems is: which information channel to use for deducing emotions? People use different modalities to express their emotions, such as facial region, body language and human speech. According to Mehrabian [8], there are three basic elements that can be used during any face-to-face communication: (1) face, (2) speech acoustics, (3) spoken words. Facial expression and speech acoustics are the most frequent informatics channels with 55% and 38%, respectively, while spoken words channel is the least contributing one with 7% of the overall impression. However, combination of these channels can increase the level of information existing to deduce the emotion state. Thus, a wide range of methods for combining various modalities with different levels of fusion have been developed [9].

Different survey papers have been published within emotion recognition task for social robots. In [2], the authors classified emotional HRI systems according to the basic stages of emotion recognition methodology (*i.e.*, sensing modalities, feature extraction and learning models). However, few articles were focused on their works on defining the challenges of building such HRI systems and describing possible sensing modalities in details [1]. For the analysis purpose, references used in this survey have been extracted from the IEEE Xplore, Google Scholar and Research Gate databases by using the following keywords: human robot interaction, HRI applications, emotion theory, uni-model emotion recognition, multimodal emotion recognition, and data fusion. We collected 213 papers published over a span of seven years from 2012 to 2019 about the aspects of recognizing, interpreting, and implementing emotions in HRI systems. However, we excluded the papers concerning the hardware implementation of emotions in robotics (10 papers) and focused on the papers containing detailed description of the methodologies used with complete experimental results and full analyses.

Organization of the remaining sections of this article is as follows. We start with illustrating existing emotion theories in section 2. Section 3 outlines the challenges of building emotion recognition systems for HRI. Section 4 reviews possible information channels used to deduce the emotion state with a review of recent works published in each channel. Finally, issues and recommendations for future works are presented in Section 5.

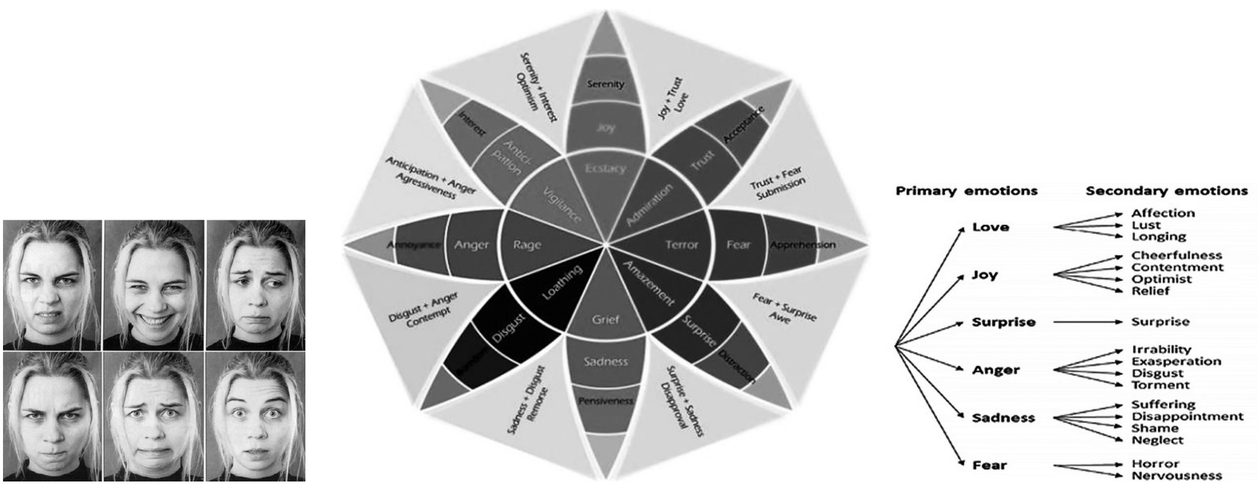
2. Emotion Models

Emotion is a complex human function that can be discussed from physiological, cognitive, and motivational points of view. A psychologist would definitely give a different definition for the term emotion from that of a linguist, a computer scientist or a normal person [10]. According to the researches in psychology, there are three theoretical perspectives on emotion: *categorical*, *dimensional and appraisal-based* *theories* [11]. Categorical emotion theory is based on a discrete set of basic emotions (primary emotions) and, in some cases, of secondary emotions generated as combinations of the primary ones. The most famous and widely accepted approach to basic emotions was conducted by Ekman. In this model, according to Ekman [10], a set of six basic emotions can be defined: joy, disgust, anger, fear, surprise and sadness, as depicted in Figure 2(a). Another categorical emotion model is proposed by Plutcnik. Plutcnik preferred to present his model with eight basic emotions by a wheel analogous to the well-known color wheel, as depicted in Figure 2(b) [12]. In Plutchik's model, the eight basic emotions are presented in opposite pairs (anger vs. fear, anticipation vs. surprise, trust vs. disgust, joy vs. sadness) on this wheel. The distance of the position of each emotion from the center of the wheel models the activation of the corresponding emotion [13]. In addition to Ekman's and Plutchik's models, Parrott's model also gained notoriety in categorical emotion theory. Parrott identified over 100 emotions and visualised them as a tree-structured list, as seen from Figure 2(c). The first layer is composed of six primary emotions (love, joy, surprise, anger, sadness, and fear) that can be branched out into different forms of feeling, and the secondary emotions are a derivation of the primary ones instead of being a combination of them [14].

In contrast to categorical emotion theory, the dimensional emotion theory argues that emotions are not independent but are related to each other in a systematic way and can be represented in a common multidimensional space [15]. Employing this theory, the raters describe different verbal stimuli on bipolar scales consisting of two opposite adjective pairs, exp., hot-cold, white-black, fast-slow, *etc.* The two primary dimensions are *valence*, which refers to how positive or negative an emotion is, and *arousal*, which describes the intensity of an emotion (ranging from sleepiness to excitement) [16].

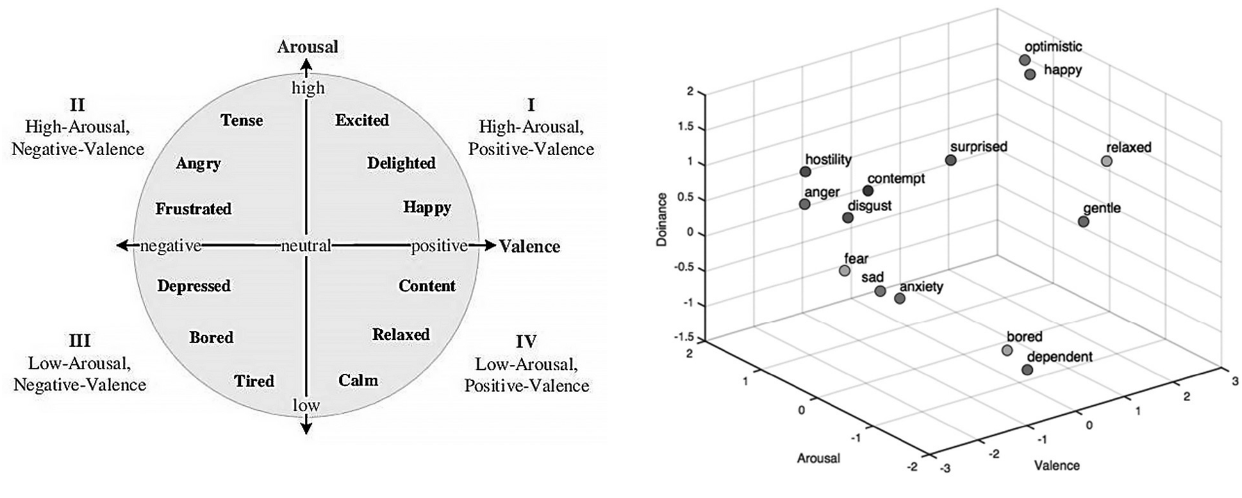
In addition to the two-dimensional representations, the three-dimensional model of emotion presented by Bradley and Lang is also used. In this model, they labeled the dimensions as: *valence*, which ranges from negative to positive emotion, *arousal*, which ranges from calm to highly aroused, and *dominance*, which describes if the person is controlled by or controlling the emotion [17]. Figure 3 shows some emotions located in both two- and three-dimensional spaces according to the above mentioned models.

The appraisal-based emotion theory claims that emotions arise from one's perceptions and cognitive evaluations of their circumstances. The theory accounts for individual variances of emotional reactions to the same event. However, its application to automatic emotion recognition is still in the early stages [18].



(a) *(b)* *(c)*

Figure 2. Three well-known models of categorical emotion theory, *(a)* basic emotions of Ekman's model, *(b)* color-wheel-like of Plutchik's emotion model, *(c)* tree-structured like Parrott's model.



(a) (b)

Figure 3. The dimensional emotion theory, *(a)* Two-dimensional valence-arousal model, *(b)* 3D emotions space model.

3. Challenges of Building Emotion Recognition System For HRI

We are entering an amazing new era full of feelings, autonomy, and computer reasoning. Thus, emotional robots are created to a dimension where they can communicate in social environments.

However, several difficulties and research issues appear during building systems for the emotional robots. The researchers should consider the following challenges during building robotic systems with emotion recognition ability.

Changing environments. Emotional robots must be placed in real world environments and they must have the ability of interaction with changed settings in a dynamic manner. Factors such as occlusion, uneven lighting, noisy sound, complex background, low resolution camera or poor microphone may affect the accuracy of the emotion detection ability of the robot. The researchers must use powerful features to create new models and they may utilize more biologically inspired techniques that can adapt to changing environmental conditions [19].

Level of description. With little exclusion, a small set of emotional states are carried out in the experiments of most developed emotion recognition systems (*i.e.*, disgust, fear, surprise, happy, sad, anger, and neutral) [20]. However, in everyday life, there are many other emotion states such as the emotions defined with Parrot model [21].

Pure and blended emotions. Mixed impressions are more complicated and more difficult to analyze than pure impressions; they are not identical and produce several expressions at the same time. In the field of emotions classification, most of the approaches are dealing with pure emotions [22] [23].

Spontaneous and acted emotions. Emotion expressions can also be classified into either spontaneous emotions or acted emotions. Saying which is happening at the moment of emotion assessment is a difficult task. People control their emotion for many reasons (such as display rules for cultural or group) and this will control how people in public manage their emotions [21][24].

The intensity of expressions. There are different variations in the intensity of the expressed emotion because under different circumstances, people can state their emotions in different ways [25].

Cross-Cultural Variations. Variations in ethnicity, national and regional backgrounds, race, in-group versus out-group relations, facial display rules within a culture and social attitudes can influence emotion recognition task [26].

Emotions of a talking face. It is difficult to describe facial emotions during talking because some muscles of the face are moved when a person is speaking and this will cause change in the appearance of face; and consequently the assessment task for expressed facial emotions becomes more difficult [22].

Different modalities to interact with robots. The investigation of new kinds of connections (modes) among people and robots can be progressively associated with people to capture the emotion state. Thus, novel sorts of association among robots and people can be developed [7].

Development of new embodies of robotics. Creating new materials and mechanical designing advancements will encourage novel encapsulations of applied autonomy. As the new embodies of robotics become available to scientists, a specialist should consolidate them with new software models of emotions, new types of expressing feelings and new use of articulations [7].

4. Emotion Recognition Sensing Channels

HRI studies can be classified into non-verbal and verbal interactive communications [27]. Verbal information involves spoken words, while the non-verbal information includes information rather than spoken words such as facial image, voice, body language. Ekman [10] proposed that non-verbal practices are the essential vehicles for communicating feelings. This is confirmed by Mehrabian [8], who found that the dominating type of correspondence is non-verbal with manner of speaking and body language representing 38% and 55% of affective information, respectively, and spoken words representing the remaining 7% (the "38% – 55% – 7% rule"). By this way, examination of non-verbal sensing channels is the key part in understanding and synthesis of emotions in robots [27].

The design of social robots searches for creating effective models for an affective interaction. In this specific situation, greater parts of the current HRI methodologies utilize a single channel of information called "mode", for example, facial expression, speech or body language [28]. Methodologies that, for example, utilize just visual or speech data separately usually fail in genuine situations. For example, occlusions, light conditions, shadows and many other factors are typical conditions where the precision of the HRI systems decreases for visual frameworks. Also, in the speech frameworks, the environmental noise or person moving while at the same time talking are considered as sound sources of mistake. To solve the problems of a single modality, emotions can be communicated through more than one modality at the same time. In multimodal emotion-based interaction, the user can transmit his emotional state to the robot utilizing diverse information channels [29]. As opposed to single modality, where each channel presents corresponding data in the emotion recognition system, a greater part of the multimodal frameworks utilizes these channels as redundant data. This redundant data is valuable in genuine interaction conditions, where, for example, mistakes related to noisy sound or partly occluded face can be diminished [30].

Thus, for an effective HRI, many authors have concentrated on multi-modal frameworks that determine the user emotion from various data sources, for example, face, motion, text, physiological signals, and many others [30]. In this paper, we classify the emotion recognition systems into two broad categories: (1) *uni-modal* and (2) *multi-modal systems* according to the number of information channels used.

4.1. Uni-Modal Emotion Recognition

Many related works on affective computing didn't combine distinctive sensing channels into a single framework for the recognition of human emotion state; instead, diverse sources of data (speech, facial expressions, bio-signals ..., *etc.*) are generally used independently to each other. The following subsections describe different sensing channels that can be utilized to detect human emotions [31].

4.1.1. Facial Expressions

Facial expressions are among the fundamental data in any face-to-face communication. Thus, it is natural that exploration of facial impressions has gained a lot of interest many decades ago, with diverse applications in effective computing and HRI [32].

Facial emotional recognition is essentially a pattern recognition problem and involves finding an optimal feature set from the facial data being analyzed [33]. Various techniques are followed to carry out feature extraction task which can fall basically into the following two main categories [34].

Appearance-based features. The appearance-based features appear temporarily in the face during any kind of facial expression (for example, the presence of specific facial wrinkles, bulges, forefront and the texture of the facial skin in the regions surrounding the mouth and eyes). Transform filters, such as Haar wavelets and integral image filters, are applied to these regions to extract the discriminative feature vector [35]. Mittal *et al.* (2012) [36] utilized Principal Component Analysis (PCA) and Euclidian distance measurement to identify the expressed emotion and then control the directions of the robot movements. The experiments on a database consisting of 150 images differing in light conditions and background complexity showed that the accuracy of the proposed system was 97.3%. Banda *et al.* (2015) [37] used LBP-Top algorithm for texture feature extraction, PCA, Locality Preserving Projections (LPP) and Factor Analysis (FA) for feature selection. Final feature set was classified using Nonlinear Auto Regressive with eXogenous inputs Recurrent Neural Networks (NARX-RNN) with an accuracy equal to 95%. Liu *et al.* (2017) [38] utilized 2D-Gabor, Local Binary Pattern (LBP), and multiclass Extreme Learning Machine (ELM) classifier, which are applied to real-time facial expression recognition for robots to recognize the seven basic emotions. The system showed the accuracy equal to 80% on the Japanese Female Facial Expression (JAFFE) public dataset. Mohammed and George (2017) [39] extracted texture features from face region by measuring the energy of the facial image blocks after applying Haar wavelets transform. Feed forward neural network classifier is then used for classification task with the accuracy equal to 90.05% using JAFFE dataset as test material. However, appearance-based features are not robust against different facial variations (like scaling of the face, appeared face region, and head area orientation *etc.*) and cannot work well with noisy images [40] [41].

Geometric-based features. In the geometric method, location of key facial components such as eyes, eyebrows, mouth and nose are being tracked and the geometric relationship between certain key points (*i.e.*, fiducial points) on the face (*e.g.*, distances, angles and shapes) are considered when making the decision [42]. Using analytical geometry calculations, Mohammed and George (2015) [43] measured the geometric feature (distances and angles) set from the basic facial components such as eyes, eyebrows and mouth. JAFFE have been used as test material and the achieved accuracy was 95.73%. Meghdari *et al.* (2016) [44] developed a vision system for the robot to identify emotional state of the user based on the facial points detected by the vision system used. A state machine was then used to identify the user's emotional state and generate the robot's reaction. An accuracy of about 90% was achieved on a database consisting of 3000 samples from 10 persons obtained using the Kinect sensor. Geometric-based features are robust against different facial variations like scaling of the face, face position within the scene, appeared face region, and head area orientation [45]. However, the extraction of such type of features can be considered expensive from the computational point of view, because it requires reliable and accurate methods for facial feature tracing and detection [46].

Several attempts have been reported using both geometric and appearances based features to overcome the limitations of both types. Such methods are referred to as "hybrid methods". In general, all these methods have proved difficult to extract the feature vector that can accurately distinguish the emotion type from facial image. Thus, the researchers attempted to use other approaches rather than the two basic types to extract discriminated features. Some of these approaches are, as follows.

Using Ekman Action Units (AUs). AUs simulate facial muscle movement during the facial expression, as shown in Figure 4. The emotion state can be determined using a combination of facial action units [47]. Hsu *et al.* (2017) [48] used Gabor filter and Support Vector Machine

(SVM) to identify the AUs. After that, random forest classifier was used to recognize the emotional state based on the identified AUs. Results of the experiments on Cohen-Kanade+ (CK+) database showed that the system can identify the facial emotions with accuracy of about 95%. However, AUs recognition is a challenging problem due to different factors such as illumination changes, pose variations or individual subject differences [49].

Figure 4. Ekman Action Units (AUs).

Using deep learning techniques. With the advances on neuro-based learning algorithms, deep learning algorithms found their wide applications on facial emotion recognition task. Deep learning is the application of multi-neuron multi-layer neural networks to perform learning tasks, including classification, clustering, regression, encoding, decoding and others [50]. The deep learning techniques are classified into two main architectures: Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). CNN is a feed forward network basically used in image processing. The strength of CNN depends on the number of hidden layers used between the input and the output layers. Each layer extracts a set of features. Feature maps are generated by applying a series of filters over the input image. Each filter goes through the entire image and multiplies its weights by the input values. RNN uses back propagation concept in which, along with the current inputs, it also considers the previous inputs. RNNs can process sequential data with the help of a memory cell. RNNs are designed based on the principle that humans do not think from scratch every time [51]. Zhang (2018) [52] used CNN to deal with emotion recognition problem from facial expression. The author extracted multiple input features and used mask loss to focus on the valid local facial features to handle complex environments, such as blurry face details and bad light conditions. The Emotion Recognition in the Wild Challenge, Static Facial Expression Recognition sub-challenge (SFEW) database was used as test material and the proposed system provided an improvement of about 35.38% over baseline results. Deep learning techniques provide higher accuracy and better predictive performance, and are more flexible and configurable [50]. Despite these advances, selection of optimal features and parameters in deep learning remains a challenging issue [53].

Using graph mining techniques. Recently, the use of graph representations has gained popularity in pattern recognition and machine learning fields. The main advantage of graphs is that they offer a powerful way to represent structured data [54]. By taking graph mining as a starting point, a novel method was proposed by Hassan and Mohammed (2019) [55] for facial emotion detection by utilizing graph mining techniques to make a paradigm shift in the way of representing the face region as a graph of nodes and edges. The graph-based Substructure PAtterN (gSpan) algorithm has been used to find frequent sub-structures in the graph database of each emotion. Final system accuracy was 90.00% using Surrey Audio-Visual Expressed Emotion (SAVEE) database.

4.1.2. Speech

Human speech includes information about both attitude and content. Recognition of speech emotion, which is a part of effective computing, extracts emotional states from the discourse and uncovers the attitudes implied within the spoken language [56]. One of the primary difficulties in speech-based emotion recognition is distinguishing the emotions over various languages and the case of mixed languages. In general, individuals may use mixed words over many languages. For example, English words with words from an alternate language (culture related). Such pragmatic situations present real difficulties for human–robot interaction frameworks [57].

Although many speech characteristics were investigated in the recognition of speech emotions, the researchers did not identify the best speech characteristics for this task till now. The reason is the similarity between the extracted features for different emotions. However, there are two main groups of speech features [58], as follows.

Continuous-based features. Continuous features are important in the provision of speakers' emotional indication and, therefore, they are commonly used in speech emotion recognition. Many researchers believe that effective continuous features such as pitch and energy reflect most of an utterance's emotional content. For example, a speaker's arousal state influences the overall energy, and the duration of speech pauses. Some of the well known continuous acoustic features are energy related features, pitch, formants and timing related features [59]. Pitch is a fundamental property of the speech signal. It describes the highness and the lowness of tone in the speech. Pitch features increase with high-arousal emotions such as happiness and surprise while they decrease with low-arousal emotions such as sadness and fear [60]. In phonetics, formants essentially mean the acoustic resonance of the human vocal tract [60]. They can be extracted by finding amplitude peaks in the frequency spectrum of the speech. Timing-related features provide information about the distribution of duration-related parameters such speech rate, the ratio between voiced and unvoiced parts. Timing features increase with high-arousal emotions and decrease with low-arousal emotions [60]. Zhu *et al.* (2017) [61] extracted zero crossing rate, pitch, formant, and short term energy features from the speech signal. Deep Belief Network (DBN) and SVM are then combined for classification purpose. Experiments were conducted using a database built by the Chinese Academy of Sciences and the accuracy of 95.8% was achieved.

Spectral-based speech features. It is identified that the emotional state of an utterance affects on the distribution of the spectral energy throughout the frequency range of the speech signal. For instance, it is found that speech signal with happiness emotion has high energy at the high frequency range while the signal with sadness emotion has low energy at the same range [62]. However, the derived spectrum is often passed through a bank of band-pass filters to better leverage the spectral distribution over the audible frequency range. Spectral features are then computed from the outputs of these filters. An example of spectral-based speech features is the Mel Frequency Cepstral Coefficients (MFCCs) [60]. Chavhan *et al.* (2015) [63] suggested an automatic Speech Emotion Recognition (SER) system in which features such as: MFCCs and Mel Energy Spectrum Dynamic Coefficients (MEDCs) were extracted from the speech signal. The LIBSVM with Radial Basis Function (RBF) kernel were used for classification task. The accuracy achieved using Berlin database was 93.75%.

In spite of the fact that many research works are done and different applications are developed, speech emotion recognition remains a challenging assignment [64]. The main reasons for this are: (a) changeability of expression or the same emotion, (b) difficulty in determination which speech features are information-rich and which are poor, (c) differences in sentences, speakers, talking styles and rates, (d) possible existence of more than one emotion in the same articulation [65]. An example of speech emotion recognition difficulties, happy and anger both have common acoustic traits like pitch, amplitude, number of times their speech crosses zero pivot. In the same manner, sadness and fear have some mutual attributes. Therefore, problems occur during recognition in these two sets of emotions [64]. To solve these difficulties, the researchers attempted to use additional techniques to find the optimal solution for acoustic feature extraction task. Some of these techniques are, as follows.

Using digital image processing techniques. Image processing methods can be used with speech signal to extract the discriminated features after converting the signal into spectrogram image. Wang (2014) [66] extracted texture features from speech spectrogram image. First the spectrogram is transformed into a recognizable image using Fourier transform. Cubic curve is then used to improve the contrast of spectrogram image. Next, texture features are extracted from the spectrogram image by applying Laws' masks on the image to represent the emotion state. Finally, SVM is used as a classifier to obtain the results of the proposed system. Three databases were used in order to test the efficiency of the proposed system: (1) Berlin Emotional Speech Database (EMO-DB), (2) eNTERFACE corpus, and (3) one self-recorded database. For the three databases used, the proposed system gave accuracy ranges from 65.20% to 77.42%.

Identifying the language used in the speech. Identification of the language spoken may help in enhancing the ability of speech emotion detection. Deriche and Abo-absa (2017) [57] introduced a two stages speech-based emotion detection system that begins by identifying the language of the speech, and then a separate emotion recognition system has been built for each language (language dependent system) to determine type of emotion (neutral, happy, angry, and sad) form the dedicated language. Wavelets transform is used to extract the discriminated features and Hidden Markov Model (HMM) is then employed to track the changes of wavelets based feature vector to determine the spoken language. After language identification, a set of features is extracted from the speech, including pitch and MFCCs and then neural network is used to predicate the emotion class. The overall accuracy achieved for the proposed system was 93%.

Using intrasegmental features. Tian and Watson (2018) [67] proposed an emotion recognition system using intrasegmental features extracted from long monophthongs in a continuous speech. 36 vocal tract features and 11 glottal source features were initially extracted and an optimal subset was selected using Maximum Relevance Minimal Redundancy Backward Wrapping (MRMRBW). JL corpus was used to evaluate the system performance. A recognition accuracy of 70.5% was achieved regardless of vowel types for five different emotions.

4.1.3. Body Language

Body language incorporates various types of nonverbal indicators, for example, body posture, gestures, eye movements, hands movements, head and different parts of the body enabling people to impart an assortment of sentiments, thoughts and feelings. The internal state of an individual is communicated through different elements, for example, iris extension, direction of gaze, hands and legs position, the sitting style, body situation, and movement style [68].

Hands are the richest source of information about body language after the face [69]. For example, according to the position of hands, one can decide if an individual is straightforward (turning the hands inside towards the questioner) or deceptive (concealing the hands behind the back) [68]. Zhang and Yap (2012) [70] proposed a system for recognizing the target emotional gestures by using the Kinect sensor. Positions of the 7 joints (for example: head, right hand, left hand, right elbow, left elbow, left hip, and hip appropriate) were distinguished by utilizing OpenCV ("Open Source Computer Vision") library and distances between these extracted points were determined. Adaptive Resonance Theory (ART-2) was used to accomplish gesture-based emotion detection. Experimental results achieved on twenty-five test sets showed that all the emotional gestures are well classified with accuracy equal to 90%.

Gait of walking is another important indicator for determination of the emotion state using body language. Cui *et al.* (2015) [71] proposed a technique for recognizing emotion from natural walking. Firstly, wrist and ankle information were recorded by cell phone. The authors then applied a sliding average filter with various windows (w) to divide the gait signal into slices. From each slice, 114 features including frequency domain, time domain, power and distribution were extracted after that PCA was used to select the best feature set. For classification purpose SVM, decision tree classifier, multilayer perception, and random forest were used. The model for identifying anger-neutral-happy gave an accuracy of 85%, 78%, and 78%, respectively.

Despite the simplicity of acquiring body language data, it has been reported that gestures are culture related. The meaning of postures may change after some time, or it may even disappear. In addition, gestures differ from women to men. Women are believed to be more perceptive than men due to the concept of female intuition. There are some differences in the manner women and men display the body language.

This might be due to the impact of culture, body organization, clothes and makeup [69].

4.1.4. Bio-Signals

Besides the affective data sources mentioned above, there are an additional significant number of different methods for estimating emotion-related information. Since there are about a hundred of physiological systems of interest, the number of possible bio-signals is huge. In a broader sense, the assortment of bio-signals ranges from a visual examination of the patient up to the signs recorded from the body through the use of sensors [72]. Some of the bio-signal measures suggested to be related to emotions are, as follows.

Blood Pressure. Blood pressure can be defined as the amount of pressure the blood makes while it flows through the body. It has been observed that the blood pressure increases with negative emotions such as fear and anxiety, and decreases with relaxation [73].

Brain activity. Electrical activity of the brain can be measured by utilizing an electroencephalograph (EEG). It has been found out that EEG asymmetries over the frontal cortex during high arousal emotions (like joy, interest, and anger) are relatively greater in the left prefrontal cortex than in the right prefrontal cortex. On the other hand, it has been noticed that EEG asymmetries are relatively greater in the right prefrontal cortex during low arousal emotions (like sadness, fear and disgust) [73] [74].

Heart rate. Electrocardiogram (ECG) is a test used to measure electrical activity of human heart. Heart rate can be used to distinguish between positive and negative emotional reactions [72].

Muscle initiations. In addition, it is possible to utilize electromyography (EMG) to quantify emotion related initiations from either facial muscles or even other body muscles. Numerous investigations have recommended that negative emotions are related to high action of "corrugator supercilious", while low action of "corrugator supercilious" is related to positive emotions [74].

Skin temperature. A change in the temperature of the skin can be considered as an indicator of emotion states. The increase in skin temperature often relates to negative emotion states "*e.g.*, anger", while the decrease in skin temperature relates to positive emotion states "*e.g.* calm" [72].

Breathing rate. Breathing rate is the number of breaths a person makes within one minute. Negative emotions are related to quicker and more profound breath while positive emotions are related to slower and shallower breath [73].

Tactile information. "*Pressure mouse*" has been developed for measuring the pressure of a user touch. Based on this information it is possible to recognize user satisfaction and disappointment [74].

During the search of the literature for the state of art works on emotion recognition tasks using bio-signals, we found that EEG signals gained the lions' share on bio-signal based emotion recognition. So, most of our focus will be on the published works within EEG signal. In general, EEG features can be extracted either from the spatial domain or from the frequency domain representation of EEG signal.

With respect to frequency domain methods, Mikhail *et al.* (2013) [75] divided EEG signal into 29 windows, each of the width equal to two seconds, with the overlapping between windows equal to one second. Fast Fourier Transform (FFT) was then used to convert each window into a frequency domain. Power bands, theta, alpha and beta rhythms features were then extracted from the frequency domain. Recognition rates with SVM as a classifier were about 51% – 53% – 58% and 61% for joy-anger-fear and sadness, respectively. Liu (2014) [76] proposed a three layers scheme for single trial Electroencephalogram-based valence and arousal Emotion Recognition (EEG-ER). The purpose of the first layer was to extract a set of spectral power descriptors from the different frequency bands of EEG signal. The kernel Fisher's analysis method was then used in the second layer to select features with better classification ability than the extracted EEG spectral powers. In the final layer, SVM was used as a classifier. Performance of the proposed three layers solution with a self-collected dataset showed that the system gave a good performance. Alazrai *et al.* (2018) [77] applied the Choi-Williams TFD (CWD) on each EEG segment. Frequency domain features were extracted from the extracted CWD representations. The extracted features were then used to recognize the emotion class using subject-dependent SVM classifiers. DEAP dataset was utilized to test system performance and the achieved accuracy was 73.8%.

Within the spatial domain methods, Liu *et al.* (2018) [78] presented a scheme with a hybrid dimension reduction to reduce dimensionality of the 14 features generated from EEG signal. Maximum Relevance Minimum Redundancy (MRMR) was then applied to reorder the extracted features into max relevance and min redundancy. Further reduction for the extracted features was also done using PCA for selection of the principal components. Results achieved using DEAP dataset showed that the proposed system exceed the state of the art methods.

EEG-based emotions can be detected by combing the spatial domain, frequency domain, and time proprieties of the raw EEG signal. Li *et al.* (2017) [79] proposed a system for classifying human emotions by EEG. Different frequency based features were extracted from various channels of EEG signals and mapped to a 2-dimensional plane to build the EEG MFI. The authors used deep neural network to recognize the emotions from EEG MFI by combining CNNs and Long Short Term-Memory Recurrent Neural Networks (LSTM-RNN). Accuracy of about 75.21% was achieved using DEAP.

Bio-signals give sensuality data that is hardly possible to access with video, speech or gesture. For this reason, bio-signals are considered as a good supplement for emotion detecting systems. Additionally, bio-signal sensors are affordable and can measure various types of bio-signals simultaneously. In spite of all these favorable features, current bio-signal electrodes are still unaesthetic, uncomfortable, intrusive, and difficult to set up, which legitimizes their use. However, new wearable devices with dry electrodes are rising and this opens a new exploration area in HRI systems [80].

4.1.5. Text

Text is the most commonly used way by humans when interacting with computers. For this reason, Sentiment Analysis (SA) of text is considered as an important issue in affective computing researches [81]. Emotion detection in conversations is a necessary step for a number of applications such as argumentation mining, understanding consumer feedback in live conversations... *etc.* However, recognizing textual emotions is a major challenge for both humans and machines. It needs advanced techniques in natural language processing to develop the emotion-based models [82]. Currently, SA has become a hot topic for the researchers in the field of natural language processing [83].

Different techniques have been proposed by the researches within SA task. Shivhare and Saritha (2014) [81] developed a system based on keyword-spotting technique as well as on the rich features of ontology. Wang and Odbal (2014) [84] applied the hierarchical structure of sentences and dependency relationships and exploited segment based features. The experimental results showed that the proposed model gave accuracy of 65.12% for news contents dataset, 50.81% for Alm's translation dataset and 53.23% for blog dataset.

Haggag *et al.* (2015) [85] introduced a method for text based emotion recognition based on ontology. Text ontology was extracted using OpenNLP parser. After that, ontology matching was performed with the ontology-base that the authors had built previously. On the self-made dataset of 511 sentences the system gave the accuracy of about 85.99%. Asghar *et al.* (2017) [86] integrated cognitive based emotion theory with sentiment analysis based computational techniques to classify the emotions from a text, using Emotion Word Classifier (EWC), Emoticon Classifier (EC), Slang Classifier (SC) and Mixed-Mode Classifier (MMC) in a pipelined approach. Evaluation of the system using the news dataset, mobile/smart phones dataset, and ISEAR dataset showed that the achieved precision was 76.7%, 83%, and 73.73% for the three dataset, respectively.

As with speech and facial expression modalities, deep learning techniques found its place in textual emotion classification. Majumder *et al.* (2018) [87] presented a method based on recurrent neural networks (DialogueRNN) that keeps track of the individual emotion states throughout the conversation. IEMOCAP dataset was used to evaluate DialogueRNN and the achieved accuracy was 63.40%.

4.2. Multimodal Emotion Recognition

Multimodal emotion recognition obtains information received from more than one channel. However, it requires defining the location of fusion information received from each channel. In general, there are three types of data fusion methods usually used in data fusion field: (1) *data-level fusion method*, (2) *feature-level fusion method*, and (3) *decision-level fusion method*. Data-level fusion method can be achieved by merging more than one physical signals of similar nature (*e.g.*, two EEG signals, two videos of two cameras, *etc.*). Since data-fusion required that different modalities always have the same nature and same signal properties, this type of fusion is not feasible for multimodal fusion [74]. Figure 5 illustrates the data-level fusion method.

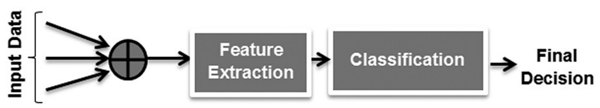


Figure 5. Data-level fusion method.

In feature-level fusion, there is a single classifier that receives all features extracted from each channel as input and makes the decision regards the user emotion using some decision algorithms. Feature-level fusion method benefits from the correlation of the extracted features in different modalities. However, it is criticized for overlooking the distinctions in temporal structure, scale and measurements among different features. Also, it requires synchronization between the modalities. It is also more difficult and computationally more expensive than performing a combination at the decision level [27]. Figure 6 shows the idea of feature-level fusion method.

Figure 6. Feature-level fusion method.

In decision-level fusion method, a classifier is assigned for each channel to make the decision about the type of recognized emotion. In this method, a choice rule must be defined to decide on the final emotion class. This rule must take into account the decision given by each channel and its certainty [27]. Decision level fusion advantages are: (1) it is simple, (2) there is no need for synchronization among modalities, and (3) it has no high computational requirements. For these reasons decision-level fusion techniques are mostly used by the researchers within the field of multimodal emotion recognition [73]. Figure 7 shows the decision-level fusion method.

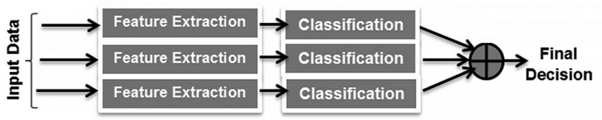


Figure 7. Decision-level fusion method.

Below are some of multimodal combinations proposed by the works published in the literature.

Face and speech modalities. Speech and facial expression are two natural and effective ways of expressing emotions when human beings communicate with each other. They have been utilized in most multimodal emotion recognition works. Prado *et al.* (2012) [88] used Ekman AUs for facial features extraction, pitch, intensity, energy and speech duration features for acoustic feature extraction. Dynamic Bayesian (DB) network was then employed for classification task in both modalities. Finally, Bayesian mixture was used at a decision level fusion. The system gave accuracy of about 82.14% when tested on 5 different emotions. Fadil *et al.* (2014) [89] exploited prosodic features in the speech and facial expressions in the videos to build a multimodal emotion recognition system. The classification of the video in one of 6 basic emotions was carried out by deep networks. The results obtained with RML emotion database were nearly 80%. Cid *et al.* (2015) [30] based on the edge image and dynamic Bayesian classifier for facial feature extraction. Dynamic Bayesian was also used with the set of speech descriptors, such as speech rate, pitch and energy. Both subsystems were combined using a third DBN classifier. The results of this multimodal approach on SAVEE database showed the robustness of the methodology with respect to single emotion recognition systems. Machine learning and optimization algorithms were combined by some authors to recognize the emotion from multimodal channels. Gharavian *et al.* (2016) [9] used audio and visual information to recognize the emotions using Fuzzy ARTMAP Neural Network (FAMNN). Particle Swarm Optimization (PSO) was utilized by the authors to find the optimal values of FAMNN parameters. The final recognition rate of the proposed multimodal system on SAVEE database reached 98.25% for the seven basic emotions. Tzirakis *et al.* (2017) [90] utilized CNN to extract features from the speech and LSTM-RNN of 50 layers was utilized in the visual modality. The experiments were conducted on the RECOLA database and the resulting accuracy was about 78.8%. Some authors used all video frames to achieve better results than using a single static image. Perez *et al.* (2018) [91] proposed the recognition of multi-modal signals to identify the four human emotions in the context of human- robot interaction. Borders and corners features were extracted from each key frame image of the video and a CNN was used for classification purpose. MFCCs features were extracted from audio data and SVM classifier was used to detect the final emotion class. A combined classifier based on CNN and SVM classifiers confidence levels was used to fuse the two classifiers results. For testing purpose a total of 110 videos were collected from 11 users using Microsoft Kinect sensor and the accuracy of about 86.4% was achieved for individual frames and 100% when used to detect emotions using all video frames while the voice classifier achieved the accuracy of about 69.7%.

Text and speech modalities. Speech and text data have been also used by some authors. Lee *et al.* (2018) [92] used CNN to extract features from the two modalities. Attention model was used to fuse these features in feature-level fusion. A third CNN was used to find the final emotion state. For evaluation purpose, CMU-MOSEI dataset was used and the achieved accuracy was about 83.11%. Gu *et al.* (2018) [56] suggested a deep multimodal framework to identify human emotions based on sentence level spoken language by extraction of the spatial features from spoken text, temporal features from speech signal. All features were fused by using deep neural network with three layers to learn the correlations among the used modalities. The proposed framework was evaluated on the IEMOCAP dataset. The result was about 60.4% for five emotion categories. Yoon *et al.* (2018) [93] proposed a deep dual recurrent encoder model that combines text and audio information at the same time to predicate the emotion class. The authors analyzed speech data from the signal-level to the language-level, and utilized the information within the data more comprehensively than it is possible with the models that focus on audio features. The proposed model is applied to the IEMOCAP dataset and the achieved accuracy equals to 71.8%.

Speech, face and gesture modalities. Besides the face and speech sensing modalities, Liu *et al.* (2016) [94] have presented a multimodal emotional communication based humans-robots interaction system (MEC-HRI) using three emotional modalities: speech, facial expression and gestures. MFCCs and other statistic features were extracted from the speech signal, after that correlation analysis methods were used for dimension reduction. Selected speech features were given to ELM model for emotion recognition. Features from the face region were extracted using Gabor filter and 2DPCA was used for dimensionality reduction. After that, another ELM model was utilized to recognize the facial expression. To extract gestures, the authors used three blobs' center average locations. At the feature level fusion, a single ELM classifier is designed for the feature vector of the three modalities. At the decision level fusion, Naïve Bayes classifier is designed for every modality. Decision probability of each basic emotion is then calculated for each modality and the emotion with the highest probability chosen as the final decision. MEC-HRI was tested on the system consisting of three NAO robots, two mobile robots, Kinect sensor, eye tracker, and two high-performance computers. The experiments showed that MEC-HRI system could recognize emotional communication between humans and robots.

In addition to the above mentioned combinations of modalities, there are many other combinations published in the literature of multimodal emotion recognition task. Table 1 demonstrates the summary of the state of art the works listed in this survey.

Table 1. Summary of the state of the art works within emotion recognition task.

Authors	Modalities	Methodology	Test Materials	Results
Liu *et al*. (2014) [76]	EEG	EEG spectral powers features, SVM for classification.	Self-collected dataset	N/A
Liu *et al.* (2018) [78]	EEG	(mRMR) for feature extraction, PCA for selecting the principal components	DEAP dataset	N/A
Li *et al.* (2017) [38]	EEG	EEG MFI for feature extraction, LSTM-RNN for classification.	DEAP database	Accuracy of about 75.21%
Alazrai *et al*. (2018) [77]	EEG	Frequency features extracted from the CWD representation of EEG segments, SVM for classification.	DEAP EEG dataset	Accuracy of about 73.8%
Mikhail *et al.* (2013) [75]	EEG	Frequency descriptors of FFT, SVM for classification.	N/A	An average accuracy of about 57%
Zhang (2018) [52]	Face	Multiple input features fusion and CNN for classification	SFEW dataset	35.38% improvement over the baseline results
Hsu *et al.* (2017) [48]	Face	Gabor filter and SVM to identify AUs, random forest classifier to recognize the emotional state.	CK+ database	Accuracy of about 95%
Banda *et al.* (2015) [37]	Face	LBP-Top algorithm for feature extraction, PCA, LPP and FA for dimension reduction, NARX-RNN for classification	SEMAINE audio-video emotional database	Accuracy of about 95%
Mittal *et al.* (2012) [36]	Face	PCA and Euclidian distance calculation	A self-made database consists of 150 images	Accuracy of about 97.3%
Mohammed and George (2017) [39]	Face	Haar wavelets transform for feature extraction, neural network for classification	JAFEE dataset	Accuracy equals 90.05%
Mohammed and George (2015) [43]	Face	Distances and angles features, neural network classifier	JAFEE dataset	Accuracy equals 95.73%
Meghdari *et al.* (2016) [44]	Face	Facial points' locations and state machine for change determination	3000 samples collected from 10 persons using Kinect sensor	Accuracy equals 90.00%

Authors	Modalities	Methodology	Test Materials	Results
Hassan and Mohammed (2019) [55]	Face	gSpan graph mining for feature extraction	SAVEE database	Accuracy equals 90.00%
Fadil *et al.* (2014) [89]	Face + Speech	Deep networks	RML emotion database	nearly 80%
Perez *et al.* (2018) [91]	Face + Speech	Borders and corners visual features and CNN for classification, MFCCs audio features	110 videos collected from 11 users using Microsoft Kinect sensor	Accuracy of about 86.4%.
Cui *et al.* (2015) [71]	Gait	114 features from time domain and frequency domain. PCA feature selection, SVM, decision tree, Multilayer perception, Random tree and random forest for classification.	Datasets of wrist and ankle	Accuracy of 85% – 78% – 78% for three emotions
Zhang and Yap (2012) [70]	Gestures	Distances between skeleton points as features and ART-2 as classifier	Dataset including 60 observations	Accuracy of about 90.00%.
Tian and Watson (2018) [67]	Speech	Intrasegmental features, MRMRBW for feature selection. DRRQM, FFNN, SVM, KNN, GMM for classification	N/A	Accuracy of about 70.5%
Deriche (2017) [57]	Speech	HMM to identify the language. MFCCs are used with NN to identify emotion type	N/A	Accuracy of about 93%
Zhu *et al.* (2017) [61]	Speech	MFCCs, pitch, formant, zero crossing rate and energy features. DBN and SVM for classification	Database created by the Chinese Academy of Sciences	Accuracy of about 95.8%
Wang (2014) [66]	Speech	Features extracted using Laws' filters on speech spectrogram image, SVM for classification	EMO-DB, eNTERFACE and KHU-SC-EmoDB	Accuracy range from 65.20% to 77.42% for the three databases, respectively
Chavhan *et al.* (2015) [63]	Speech	MFCCs and MEDC for feature extraction, LIBSVM with RBF kernel for classification	Berlin database	Accuracy equal to 93.75%
Tzirakis *et al.* (2017) [90]	Speech + Face	CNN for extracting features from the speech, deep residual network of 50 layers for extracting features from the facial region, LSTM-RNN for classification	RECOLA database	Accuracy of about 78.8%
Cid *et al.* (2015) [30]	Speech + Face	Edge features and dynamic Bayesian classifier for face image, a set of speech descriptors, dynamic Bayesian classifier for classification	SAVEE database	N/A

Authors	Modalities	Methodology	Test Materials	Results
Prado *et al.* (2012) [88]	Speech + Face	AUs and DB for facial expression, pitch, intensity, energy and speech duration features and DB for speech emotions	N/A	Accuracy of about 82.14%
Lee *et al.* (2018) [92]	Speech + Text	Two CNNs for feature extraction, attention model to fuse these features and a third CNN to find the final emotion state	CMU-MOSEI dataset	Accuracy of about 83.11%
Gharavian *et al.* (2016) [9]	Speech +Face	FAMNN and PSO for finding the optimum values of FAMNN parameters	SAVEE database	Accuracy of about 98.25%
Liu *et al.* (2016) [94]	Speech +Facial expression+ Gesture	MFCCs features from speech, Gabor filter for facial feature extraction, blobs' center average location features for gesture, ELM model for classification	A system consists of three NAO robots, two mobile robots, Kinect sensor, two high-performance computers	N/A
Majumder *et al.* (2018) [87]	Text	DialogueRNN for classification	IEMOCAP and AVEC datasets	Accuracy of about 63.40%
Asghar *et al.* (2017) [86]	Text	EWC, EC, SC and MMC emotion classifiers in a pipelined approach	News Dataset, Mobile/Smart Phones Dataset, and ISEAR Dataset	Precision of 76.7%, 83%, and 73.73% for the three datasets, respectively
Haggag *et al.* (2015) [85]	Text	Ontology matching with the ontology base	SemEval dataset	Accuracy of about 85.99%
Shivhare and Saritha (2014) [81]	Text	keyword-spotting technique and ontology features	N/A	N/A
Wang and Odbal (2014) [84]	Text	Sentences and dependency relationships and segment based features.	News contents dataset, Alm's translation dataset, blog dataset	Accuracy of 65.12% for news contents dataset, 50.81% for Alm's translation dataset and 53.23% for blog dataset
Gu *et al.* (2018) [56]	Text + Speech	Spatial features from text, temporal features from audio and deep neural network for classification	IEMOCAP dataset	Accuracy of about 60.4%
Yoon *et al.* (2018) [93]	Text + Speech	Deep dual recurrent encoder	IEMOCAP dataset	Accuracy equals 71.8%.

5. Discussion

Social robots with emotion recognition ability find many important applications within different fields in our daily life, such as in healthcare, education, entertainment..., *etc.* Even though different approaches have been proposed to develop emotion recognition systems for HRI, there are still many limitations in the developed systems. Thus, some recommendations for future works can be given, as follows.

1. Most of emotion recognition systems for HRI are built to classify the emotions listed within the categorical emotion theory [9] [30] [37–39] [43, 44] [52] [55–57] [66–67] [71] [85–86] [89–91] [93, 94]. However, HRI systems based on dimensional or arousal theories can find application in different fields. In healthcare, for example, the systems empower the robot to take care of patients according to different levels of their pain. Therefore, additional efforts are needed to develop emotion recognition system using such emotion theories.
2. Some researchers focused their work on facial expression and speech, ignoring the remaining sensing channels which may contain affective information. Recently, with the fast development of sensors' devices which can record different bio-signals, it has become feasible to build a multimodal emotional system that combines different physiological signals. This may lead to a considerable improvement in emotion recognition ability of HRI systems and giving social robots the possibility to improve people's life.
3. Most studies didn't consider dynamic environment conditions. For example, most studies trained their systems on a small number of subjects and this may lead to poor generalization and surely bad elasticity with the changed environment. The developed systems need to be tested on more realistic environments, for example in noisier environments, using more subjects of different age, from different cultures and backgrounds..., *etc.*
4. Different hybrid algorithms and complex machine learning techniques have been proposed. However, complex algorithms may take long time to execute, which makes them less efficient to be used in real time applications.
5. With the evidence of deep learning and graph mining techniques, a new research direction has been opened. The researchers may combine these two techniques to achieve more promising results within the field of emotional HRI systems.
6. Only pure emotions have been considered in the literature. However, in real life, persons can show more than one emotions (blended emotion) at the same time. For instance, a person may show surprise and happiness at the same time.
7. Within facial expressions based HRI systems, only static 2D images or dynamic 2D video sequences have been utilized. Large variations in the subtle facial behaviors are difficult to be handled using 2D image analysis. 3D facial expressions analysis can lead to better examination of the fine structural changes appearing as a consequence of the expressed emotion.
8. Most of the proposed methodologies use statistical feature selection algorithms such as PCA and Linear Discrimination Algorithm (LDA). However, different optimization algorithms such as Genetic Algorithm (GA) and PSO can be used for optimal feature set selection and this may improve the accuracy of the final system.
9. Much research has used either feature-level fusion or decision-level fusion to fuse the multimodal data. Very few works considered both types of data fusions at the same time. For example, hierarchal fusion techniques can be used by applying feature-level fusion and clustering the fused data into different clusters according to the resulting data similarities. Classifiers may be built for these clusters and then decision-level fusion can be employed.

6. Conclusion

As robots become more social, it becomes necessary to build HRI systems with emotion recognition capabilities. This paper provides a comprehensive survey on emotion recognition for HRI systems and indicates the main challenges faced by researchers when building such systems. Also, we have investigated the state of the art of different sensing channels and examined recent progress in the design of different emotion recognition systems within each modality. Finally, recommendations and trends for future work and further explorations in this area have been outlined.

References

[1] M. I. Ahmad *et al.*, "A Systematic Review of Adaptivity in Human-Robot Interaction", *Multimodal Technologies and Interact*, vol. 3, no. 14, pp. 1–25, 2017.
http://dx.doi.org/10.3390/mti1030014

[2] F. Cavallo *et al.*, "Emotion Modelling for Social Robotics Applications: A Review", *Journal of Bionic Engineering*, vol. 15, no. 2, pp. 185–203, 2018.
http://dx.doi.org/10.1007/s42235-018-0015-y

[3] L. Perez-Gaspar *et al.*, "Multimodal Emotion Recognition with Evolutionary Computation for Human-Robot Interaction", *Expert Systems with Applications*, vol. 66, pp. 42–61, 2016.
http://dx.doi.org/10.1016/j.eswa.2016.08.047

[4] T. H. Dang and A. Tapus, "Stress Game: The Role of Motivational Robotic Assistance in Reducing User's Task Stress", *International Journal of Social Robotics*, vol. 7, pp. 227–240, 2015.
http://dx.doi.org/10.1007/s12369-014-0256-9

[5] A. G. Pour *et al.*, "Human–Robot Facial Expression Reciprocal Interaction Platform: Case Studies on Children with Autism", International Journal of Social Robotics, vol. 10, no.2, pp. 179–198, 2018.

[6] M. J. Salvador *et al.*, "An Emotion Recognition Comparative Study of Autistic and Typically-Developing Children Using the Zeno Robot", in *Proc. of the IEEE International Conference on Robotics and Automation (ICRA)*, 2015, pp. 6128–6133.
http://dx.doi.org/10.1109/ICRA.2015.7140059

[7] R. A. Calvo *et al.*, "The Oxford Handbook of Affective Computing", Oxford University Press, 2015.
http://dx.doi.org/10.1093/oxfordhb/9780199942237.001.0001

[8] A. Mehrabian, "Silent Messages", Belmon, 1971.

[9] D. Gharavian *et al.*, "Audio-Visual Emotion Recognition Using FCBF Feature Selection Method and Particle Swarm Optimization for Fuzzy ARTMAP Neural Networks", Multimedia Tools and Applications, vol. 76, no. 2, pp. 2331–2352, 2016.
http://dx.doi.org/10.1007/s11042-015-3180-6

[10] P. Ekman, "An Argument for Basic Emotions", Cognition and Emotion, 1992.
http://dx.doi.org/10.1080/02699939208411068

[11] D. Grandjean *et al.*, "Conscious Emotional Experience Emerges as a Function of Multilevel, Appraisal-Driven Response Synchronization", *Consciousness and Cognition*, vol. 17, no. 2, pp. 484–495, 2008.

[12] A. Seyeditabari *et al.*, "Emotion Detection in Text: a Review", arXiv:1806.00674v1, 2018.

[13] A. Kołakowska *et al.*, "Information Systems Development and Applications", Faculty of Management University of Gdańsk, 2015.

[14] F. Nonis *et al.*, "3D Approaches and Challenges in Facial Expression Recognition Algorithms – A Literature Review", *Applied Sciences*, vol. 9, no. 3904, pp. 1–33, 2019.
http://dx.doi.org/10.3390/app9183904

[15] P. A. Kragel and K. S. LaBar, "Advancing Emotion Theory with Multivariate Pattern Classification", *Emotion Review (EMR)*, vol. 6, no. 2, 2014.
http://dx.doi.org/10.1177/1754073913512519

[16] J. Grekow, "From Content-Based Music Emotion Recognition to Emotion Maps of Musical Pieces", Springer International Publishing, 2018.
http://dx.doi.org/10.1007/978-3-319-70609-2

[17] I. Bakker *et al.*, "Pleasure, Arousal, Dominance: Mehrabian and Russell Revisited", *Current Psychology*, vol. 33, no. 3, pp. 405–421, 2014.
http://dx.doi.org/10.1007/s12144-014-9219-4

[18] S. J. Achar *et al.*, "The Psychology of Appraisal: Specific Emotions and Decision-Making", *Journal of Consumer Psychology*, vol. 25, no. 3, pp. 359–371, 2015.

[19] A. Paiva *et al.*, "Empathy in Virtual Agents and Robots: A Survey", *ACM Transactions on Interactive Intelligent Systems*, vol. 7, no. 3, pp. 1–40, 2017.
http://dx.doi.org/10.1145/2912150

[20] M. Dubey and L. Singh, "Automatic Emotion Recognition Using Facial Expression: A Review", *International Research Journal of Engineering and Technology (IRJET)*, vol. 3, no. 2, pp. 488–492, 2016.

[21] G. Zhao *et al.*, "Experiments with Facial Expression Recognition Using Spatiotemporal Local Binary Patterns", in *Proc. of the IEEE International Conference*, Beijing, 2007, pp. 1091–1094.
http://dx.doi.org/10.1109/ICME.2007.4284844

[22] G. Iatraki, "Emotional Facial Expression Recognition & Classification", Master thesis, Department of Media and Knowledge Engineering,

Delft University of Technology, The Netherlands, 2009.

[23] S. Li and W. Deng, "Deep Facial Expression Recognition: A Survey", *IEEE Transactions on Affective Computing*, pp. 1–25, 2020. http://dx.doi.org/10.1109/TAFFC.2020.2981446

[24] I. M. Revina and W. R. S. Emmanuel, "A Survey on Human Face Expression Recognition Techniques", *Journal of King Saud University Computer and Information Sciences*, 2018. http://dx.doi.org/10.1016/j.jksuci.2018.09.002

[25] A. Savran *et al.*, "Regression-Based Intensity Estimation of Facial Action Units", *Image and Vision Computing*, vol. 30, no. 10, pp. 774–784, 2012. http://dx.doi.org/10.1016/j.imavis.2011.11.008

[26] M. V. Mishra *et al.*, "Cross-Cultural Emotion Recognition and Evaluation of Radboud Faces Database with an Indian Sample", *PLOSE ONE*, vol. 13, no. 10, pp. 1–19, 2018. http://dx.doi.org/10.1371/journal.pone.0203959

[27] F. Alonso-Martín *et al.*, "A Multimodal Emotion Detection System During Human-Robot Interaction", *Sensors*, vol. 13, no.11, pp. 15549–15581, 2013. http://dx.doi.org/10.3390/s131115549

[28] D. Kulic and E. A. Croft, "Affective State Estimation for Human–Robot Interaction", *IEEE Transactions on Robotics*, vol. 23, no. 5, pp. 991–1000, 2007. http://dx.doi.org/10.1109/TRO.2007.904899

[29] E. Martínez and A. P. D. Pobil, "Object Detection and Recognition for Assistive Robots", *IEEE Robotics & Automation Magazine*, vol. 99, 2017.

[30] F. Cid *et al.*, "A Novel Multimodal Emotion Recognition Approach for Affective Human Robot Interaction", *IEEE/RSJ International Conference on Intelligent Robots and Systems*, 2015.

[31] H. Gunes and M. Pantic, "Automatic, Dimensional and Continuous Emotion Recognition", *International Journal of Synthetic Emotions*, vol. 1, no. 1, pp. 68–99, 2010. http://dx.doi.org/10.4018/jse.2010101605

[32] B. C. Cuo *et al.*, "A Brief Review of Facial Emotion Recognition Based on Visual Information", *Sensors*, vol. 18, no. 2, 2018.

[33] D. Mehta *et al.*, "Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality", *Sensors*, vol. 18, pp. 1–24, 2018.

[34] C. Wu *et al.*, "Survey on Audiovisual Emotion Recognition: Databases, Features, and Data Fusion Strategies", *APSIPA Transactions on Signal and Information Processing*, vol. 3, pp. 1–18, 2014. http://dx.doi.org/10.1017/ATSIP.2014.11

[35] D. P. Tian, "A Review on Image Feature Extraction and Representation Techniques", *International Journal of Multimedia and Ubiquitous Engineering*, vol. 8, no. 4, pp. 385–396, 2013.

[36] R. Mittal *et al.*, "Autonomous Robot Control Using Facial Expressions", *International Journal of Computer Theory and Engineering*, vol. 4, no. 4, pp. 631–635, 2012. http://dx.doi.org/10.7763/IJCTE.2012.V4.546

[37] N. Banda *et al.*, "Feature Reduction for Dimensional Emotion Recognition in Human-Robot Interaction", *IEEE Symposium Series on Computational Intelligence*, 2015. http://dx.doi.org/10.1109/SSCI.2015.119

[38] Z. Liu *et al.*, "A Facial Expression Emotion Recognition Based Human-Robot Interaction System", *IEEE/CAA Journal of Automatica Sinica*, vol. 4, no. 4, pp. 668–676, 2017. http://dx.doi.org/10.1109/JAS.2017.7510622

[39] S. N. Mohammed and L. E. George, " Emotion Recognition Based on Texture Analysis of Facial Expressions Using Wavelets Transform", *Australian Journal of Basic and Applied Sciences*, vol. 1, no. 5, pp. 1–11, 2017.

[40] V. Verma and L. K. Sharma, "A Comprehensive Survey on Human Facial Expression Detection", *International Journal of Image Processing (IJIP)*, vol. 7, no. 2, pp. 171–182, 2013.

[41] S. Dhall and P. Sethi, "Geometric and Appearance Feature Analysis for Facial Expression Recognition", *International Journal of Advanced Engineering Technology*, vol. 7, no. 111, pp. 1–11, 2014.

[42] O. Celiktutan *et al.*, "A Comparative Study of Face Landmarking Techniques", *EURASIP Journal on Image and Video Processing*, vol. 13, no. 1, pp. 1–27, 2013. http://dx.doi.org/10.1186/1687-5281-2013-13

[43] S. N. Mohammed and L. E. George, "Subject Independent Facial Emotion Classification Using Geometric-Based Features", *Applied Sciences, Engineering and Technology*, vol. 11, no. 9, pp. 1030–1035, 2015. http://dx.doi.org/10.19026/rjaset.11.2145

[44] A. Meghdari *et al.*, "Spontaneous Human-Robot Emotional Interaction through Facial Expressions", in *Proc. of the International Conference on Social Robotics*, 2016, pp. 351–361. http://dx.doi.org/10.1007/978-3-319-47437-3_34

[45] S. Dutta and V. B. Baru, "Review of Facial Expression Recognition System and Used Datasets", *International Journal of Research in Engineering and Technology (IJRET)*, vol. 2, no. 12, pp.641–645, 2013.

[46] C. P. Sumathi *et al.*, "Automatic Facial Expression Analysis: A Survey", *International Jour-*

nal of Computer Science & Engineering Survey (IJCSES), vol. 3, no. 6, pp. 47–59, 2012.

[47] M. Ghayoumi and A. K. Bansal, "Unifying Geometric Features and Facial Action Units for Improved Performance of Facial Expression Analysis", *New Developments in Circuits, Systems, Signal Processing, Communications and Computers*, 2015, pp. 259–266.

[48] S. Hsu *et al.*, "Facial Expression Recognition for Human-Robot Interaction", in Proc. of the *2017 1st IEEE International Conference on Robotic Computing (IRC)*, 2017, pp. 1–7. http://dx.doi.org/10.1109/IRC.2017.12

[49] A. Ruiz *et al.*, "From Emotions to Action Units with Hidden and Semi-Hidden-Task Learning", in *Proc. of the 2015 IEEE International Conference on Computer Vision (ICCV)*, 2015, pp. 13703–3711. http://dx.doi.org/10.1109/ICCV.2015.422

[50] W. G. Hatcher and W. Yu, "A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends", *Human-Centered Smart Systems and Technologies*, vol. 6, pp. 24411–24432, 2018. http://dx.doi.org/10.1109/ACCESS.2018.2830661

[51] M. I. Prabha and G. U. Srikanth, "Survey of Sentiment Analysis Using Deep Learning Techniques", in *Proc. of the 2019 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 2019, pp. 1–9. http://dx.doi.org/10.1109/ICIICT1.2019.8741438

[52] Z. Zhang, "Deep Face Emotion Recognition", *Journal of Physics Conference Series*, vol. 1087, no. 6, 2018. http://dx.doi.org/10.1088/1742-6596/1087/6/062036

[53] O. Sharma, "Deep Challenges Associated with Deep Learning," in *Proc. of the 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon)*, 2019, pp. 72–75. http://dx.doi.org/10.1109/COMITCon.2019.8862453

[54] D. J. Cook and L. B. Holder, "Mining Graph Data", John Wiley and Sons, 2006. http://dx.doi.org/10.1002/0470073047

[55] A. K. Hassan and S. N. Mohammed, "A Novel Facial Emotion Recognition Scheme Based on Graph Mining", Defence Technology, 2019. http://dx.doi.org/10.1016/j.dt.2019.12.006

[56] Y. Gu *et al.*, "Deep Multimodal Learning For Emotion Recognition in Spoken Language", *IEEE 2018 International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2018)*, 2018, pp. 5079–5083.

[57] M. Deriche and A. H. Abo-absa, "A Two-Stage Hierarchical Bilingual Emotion Recognition System Using a Hidden Markov Model and Neural Networks", Arabian Journal for Science and Engineering, vol. 42, no. 12, pp. 5231–5249, 2017. http://dx.doi.org/10.1007/s13369-017-2742-5

[58] S. Bhadra *et al.*, "Study on Feature Extraction of Speech Emotion Recognition", *ADBU-Journal of Engineering Technology*, vol. 4, no. 1, pp. 7–9, 2016.

[59] K. Wang *et al.*, "Speech Emotion Recognition Using Fourier Parameters", *IEEE Transactions On Affective Computing*, vol. 6, no. 1, pp. 69–75, 2015.

[60] S. Sondhi *et al.*, "Acoustic Analysis of Speech under Stress", *International Journal of Bioinformatics Research and Applications*, vol. 11, no. 5, pp. 417–432, 2015. http://dx.doi.org/10.1504/IJBRA.2015.071942

[61] L. Zhu *et al.*, "Emotion Recognition from Chinese Speech for Smart Affective Services Using a Combination of SVM and DBN", *Sensors*, vol. 17, no. 7, pp. 1694–1708, 2017. http://dx.doi.org/10.3390/s17071694

[62] D. Bitouk *et al.*, "Class-Level Spectral Features for Emotion Recognition", *Speech Commun*, vol. 52, pp. 613–625, 2010. http://dx.doi.org/10.1016/j.specom.2010.02.010

[63] D. Chavhan *et al.*, "Speech Emotion Recognition Using RBF Kernel of LIBSVM", in *Proc. of the 2015 2nd International Conference on Electronics and Communication Systems (ICECS)*, 2015, pp. 1132–1135. http://dx.doi.org/10.1109/ECS.2015.7124760

[64] M. Pervaiz and T. A. Khan, "Emotion Recognition from Speech using Prosodic and Linguistic Features", *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 7, pp. 84–90, 2016.

[65] H. Muthusamy *et al.*, "Particle Swarm Optimization Based Feature Enhancement and Feature Selection for Improved Emotion Recognition in Speech and Glottal Signals", *PLOS ONE*, vol. 10, no. 3, pp. 1–20, 2015. http://dx.doi.org/10.1371/journal.pone.0120344

[66] K. Wang, "The Feature Extraction Based on Texture Image Information for Emotion Sensing in Speech", *Sensors*, vol. 14, no. 9, pp. 16692–16714, 2014. http://dx.doi.org/10.3390/s140916692

[67] L. Tian and C. I. Watson, "Emotion Recognition Using Intrasegmental Features of Continuous Speech", in *Proc. of the 17th International Australasian Conference on Speech Science and Technology*, 2018, pp. 61–64.

[68] F. Noroozi *et al.*, "Survey on Emotional Body Gesture Recognition", *IEEE Transactions on Affective Computing*, pp. 1–99, 2018. http://dx.doi.org/10.1109/TAFFC.2018.2874986

[69] P. Molchanov *et al.*, "Hand Gesture Recognition with 3D Convolutional Neural Networks", in *Proc. of the 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2015, pp. 1–7.
http://dx.doi.org/10.1109/CVPRW.2015.7301342

[70] L. Zhang and B. Yap, "Affect Detection from Text-Based Virtual Improvisation and Emotional Gesture Recognition", *Advances in Human-Computer Interaction*, pp. 1–12, 2012.
http://dx.doi.org/10.1155/2012/461247

[71] L. Cui *et al.*, "Emotion Detection from Natural Walking", International Conference on Human Centered Computing, 2015, pp. 23–33.
http://dx.doi.org/10.1007/978-3-319-31854-7_3

[72] R. K. Nawasalkar *et al.*, "Study of Comparison of Human Bio-Signals for Emotion Detection Using HCI", *International Journal of Emerging Trends & Technology in Computer Sciences*, vol. 2, no. 2, pp. 449–452, 2013.

[73] M. T. Quazi, "Human Emotion Recognition Using Smart Sensors", PhD Dissertation in Electrons and Communication Engineering, Massey University, New Zealand, 2012.

[74] A. Rabie, "Audio-Visual Emotion Recognition for Natural Human-Robot Interaction", PhD Dissertation, Bielefeld University, 2010.

[75] M. Mikhail *et al.*, "Using Minimal Number of Electrodes for Emotion Detection Using Brain Signals Produced From a New Elicitation Technique", International Journal of Autonomous and Adaptive Communications Systems, vol. 6, no. 1, pp. 80–97, 2013.
http://dx.doi.org/10.1504/IJAACS.2013.050696

[76] Y. Liu *et al.*, "Emotion Recognition from Single-Trial EEG Based on Kernel Fisher's Emotion Pattern and Imbalanced Quasiconformal Kernel Support Vector Machine", *Sensor*, vol. 14, no. 8, pp. 13361–13388, 2014.
http://dx.doi.org/10.3390/s140813361

[77] R. Alazrai *et al.*, "EEG-Based Emotion Recognition Using Quadratic Time-Frequency Distribution", *Sensors*, vol. 18, no. 8, 2018.
http://dx.doi.org/10.3390/s18082739

[78] J. Liu *et al.*, "Emotion Detection from EEG Recordings Based on Supervised and Unsupervised Dimension Reduction", Concurrency and Computation Practice and Experience, vol. 30, no. 23, pp. 1–13, 2018.
http://dx.doi.org/10.1002/cpe.4446

[79] Y. Li *et al.*, "Human Emotion Recognition with Electroencephalographic Multidimensional Features by Hybrid Deep Neural Networks", *Applied Sciences*, vol. 7, no. 10, 2017.
http://dx.doi.org/10.3390/app7101060

[80] A. Cruz *et al.*, "Facial Expression Recognition Based on EoG Toward Emotion Detection for Human-Robot Interaction", in *Proc. of the International Conference on Bio-inspired Systems and Signal Processing*, 2015, pp. 31–37.
http://dx.doi.org/10.5220/0005187200310037

[81] S. N. Shivhare and S. K. Saritha, "Emotion Detection from Text Documents", *International Journal of Data Mining & Knowledge Management Process*, vol. 4, no. 6, pp. 51–57, 2014.
http://dx.doi.org/10.5121/ijdkp.2014.4605

[82] H. S. Jayswal *et al.*, "A Review on Emotion Detection and Recognition from Text", *Conference on Natural Language Processing*, 2018.

[83] C. N. D. Santos and M. Gatti, "Deep Convolutional Neural Networks for Sentiment Analysis of Short Texts", in *Proc. of the COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, 2014, pp. 69–78.

[84] Z. Wang and Odbal, "Segment-Based Fine-grained Emotion Detection for Chinese Text", in *Proc. of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing*, 2014, pp. 52–601.

[85] M. Haggag *et al.*, "Ontology-Based Textual Emotion Detection", *International Journal of Advanced Computer Science and Applications*, vol. 6, no. 9, pp. 239–246, 2015.
http://dx.doi.org/10.14569/IJACSA.2015.060932

[86] M. Z. Asghar *et al.*, "Sentence-Level Emotion Detection Framework Using Rule-Based Classification", *Cognitive Computation*, vol. 9, no. 6, pp. 868–894, 2017.
http://dx.doi.org/10.1007/s12559-017-9503-3

[87] N. Majumder *et al.*, "DialogueRNN: An Attentive RNN for Emotion Detection in Conversations", *Association for the Advancement of Artificial Intelligence*, vol. 33, 2018.
http://dx.doi.org/10.1609/aaai.v33i01.33016818

[88] J. A. Prado *et al.*, "Visuo-Auditory Multimodal Emotional Structure to Improve Human-Robot-Interaction", *International Journal of Social Robotics*, vol. 4, no. 1, pp. 29–51, 2012.
http://dx.doi.org/10.1007/s12369-011-0134-7

[89] C. Fadil *et al.*, "Multimodal Emotion Recognition Using Deep Networks", *VI Latin American Congress on Biomedical Engineering CLAIB 2014*, 2014, pp. 813-816.
http://dx.doi.org/10.1007/978-3-319-13117-7_207

[90] P. Tzirakis *et al.*, "End-to-End Multimodal Emotion Recognition Using Deep Neural Networks", *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 8, pp. 1301–1309, 2017.
http://dx.doi.org/10.1109/JSTSP.2017.2764438

[91] A. K. Perez *et al.*, "Identification of Multimodal Signals for Emotion Recognition in the Context of Human-Robot Interaction", *International Sym-*

posium on Intelligent Computing Systems, 2018, pp. 67–80. http://dx.doi.org/10.1007/978-3-319-76261-6_6

[92] C. Lee *et al.*, "Convolutional Attention Networks for Multimodal Emotion Recognition from Speech and Text Data", *Challenge-HML*, pp. 28–34, 2018.

[93] S. Yoon *et al.*, "Multimodal Speech Emotion Recognition Using Audio and Text", *2018 IEEE Spoken Language Technology Workshop (SLT)*, 2018, pp. 112–118. http://dx.doi.org/10.1109/SLT.2018.8639583

[94] Z. T. Liu *et al.*, "A Multimodal Emotional Communication Based Humans-Robots Interaction System", in *Proc. of the 35th Chinese Control Conference*, 2016, pp. 6363–6368. http://dx.doi.org/10.1109/ChiCC.2016.7554357

Received: July 2019
Revised: February 2020
Accepted: April 2020

Contact addresses:
Suhaila Najim Mohammed
Department of Computer Science
University of Technology
Baghdad
Iraq
e-mail: suhailan.mo@sc.uobaghdad.edu.iq

Alia Karim Abdul Hassan
Department of Computer Science
University of Technology
Baghdad
Iraq
e-mail: 110018@uotechnology.edu.iq

Suhaila Najim Mohammed received her BSc degree (2010) in computer science from Baghdad University, Iraq. She received the MSc degree (2016) in computer science from the University of Baghdad, Iraq. In 2011, she started to work as a lecturer in Computer Science Department at the College of Science, University of Baghdad, Iraq. Currently, she is working toward the PhD degree at the University of Technology, Iraq. Her current research topics are pattern recognition, multimedia, data mining, soft computing and artificial intelligence.

Alia Karim Abdul Hassan was born on March 28, 1971. She received her BSc degree (1993), MSc degree (1999) and the PhD degree (2004) from Computer Science Department at the University of Technology, Baghdad, Iraq. Since 2007, dr. Abdul Hassan has supervised 26 MSc and PhD theses in computer science and published more than 85 papers in local and international conferences and journals, including Springer, IEEE, and Elsevier. She has held the position of Dean of Computer Science Department at the University of Technology since Feb. 2019. Her current research interests are: soft computing, green computing, AI, data mining, software engineering, electronic management, and computer security.