The Information Channels of Emotion Recognition: A Review

Ahmed Samit Hatem
Department of Software, Faculty of Information Technology, University of Babylon, Babylon, Iraq.
Department of Basic Sciences, Faculty of Nursing, University of Kerbala, Kerbala, Iraq.
Email: ahmed.samit@uokerbala.edu.iq

Abbas M. Al-Bakry
University of Information Technology and Communication, Baghdad, Iraq.
E-mail: abbasm.albakry@uoitc.edu.iq

Received August 18, 2021; Accepted December 02, 2021
ISSN: 1735-188X
DOI: 10.14704/WEB/V19I1/WEB19064

Abstract

Humans are emotional beings. When we express about emotions, we frequently use several modalities, whether we want to so overtly (i.e., Speech, facial expressions,..) or implicitly (i.e., body language, text,..). Emotion recognition has lately piqued the interest of many researchers, and various techniques have been studied. A review on emotion recognition is given in this article. The survey seeks single and multiple source of data or information channels that may be utilized to identify emotions and includes a literature analysis on current studies published to each information channel, as well as the techniques employed and the findings obtained. Ultimately, some of the present emotion recognition problems and future work recommendations have been mentioned.

Keywords

Emotion Recognition, Facial Expression, Speech, Multi-channels.

Introduction

Emotion is important in many aspects of our life and may impact or perhaps even determine our thinking and decision-making [Zhao *et al*., 2021]. When designing emotion detection systems, one essential question to consider is: which information channels should be employed to deduce emotions? People communicate their emotions through many modes like human speech, body language, and facial expression. Mehrabian [Mehrabian, 1971] states that three essential factors may be employed throughout any face-to-face interaction: the facial, speech acoustics, and the spoken words. Voice acoustics and face expression are the widely significant channels, accounting for 38% and 55% of the entire impression, respectively, whereas spoken words account for just 7% of such overall impression. Therefore, combining these channels can enhance the amount of information available to determine the emotional state. As a result, a variety of strategies for merging distinct modalities with varying degrees of fusion have been devised [Gharavian, Bejani and Sheikhan, 2016]. Several survey articles on emotion recognition tasks from single or several Modalities have been published. The authors of [Gharavian, Bejani and Sheikhan, 2016] detailed the emotion databases, pre-processing procedures, different feature extraction techniques, feature selection approaches, and classifiers utilized in voice emotion identification as single Modalities.

Some articles, on the other hand, concentrated on providing a comprehensive guide on various aspects of Multiple modal emotion recognition, such as affective modalities, psychological models, challenges data collections, computational approaches, and applications [Zhao *et al.*, 2021]. We gathered a collection of papers published between 2012 and 2021 on issues of emotion recognition, interpretation, and implementation in single or multiple Modalities. We concentrated on papers that provided a thorough discussion of the techniques employed, as well as complete experimental findings and analysis.

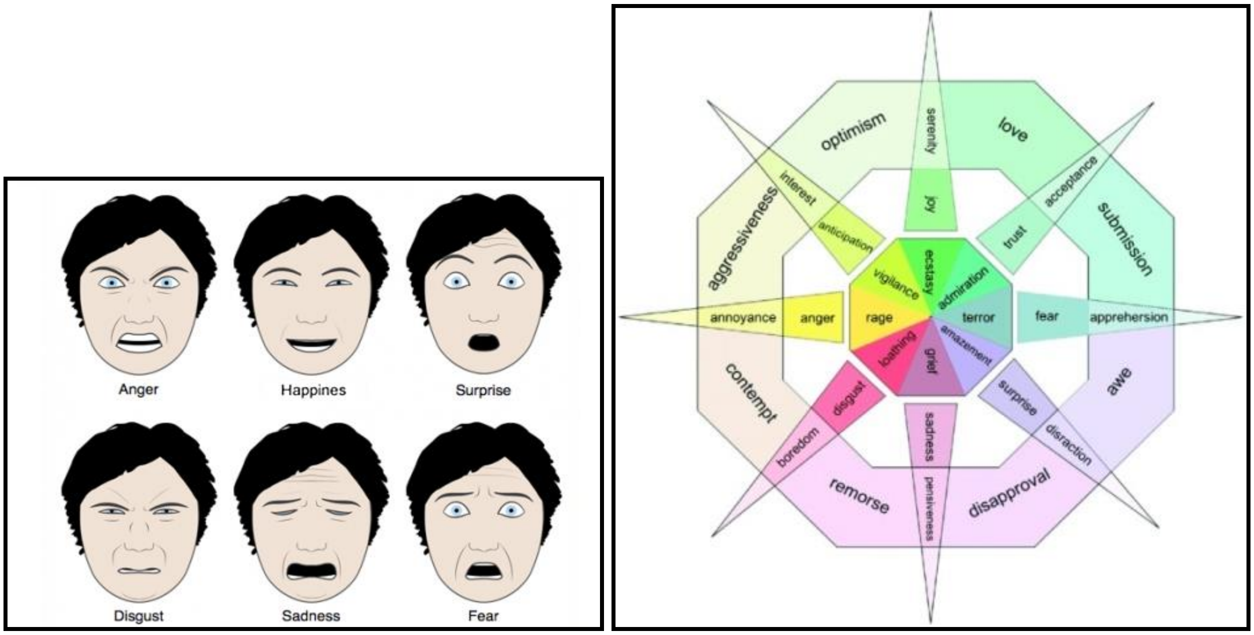
The following portions of this article are organized as follows. In section 2 we begin by demonstrating current emotion theories. Section 3 shows the single and multiple information channels that may be utilized to determine an individual's emotional state, as well as recent research in each channel. Section 4 discusses and presents recommendations for future studies. Lastly, part 5 presents the paper's conclusion.

Theories of Emotion

Emotion is complicated human process that may be studied from cognitive, physiological, and motivational perspectives. A psychologist's concept of emotion would undoubtedly differ from that of a computer scientist, linguist, or an ordinary person [Ekman,1992]. There are three main theoretical views on emotion, according to psychology studies: dimensional, categorical (basic), and appraisal theories. According to the dimensional emotion concept, emotions are not autonomous but are connected in a systematic fashion and may be depicted on a shared multidimensional space. Using this theory, raters characterize various verbal stimuli onto bipolar scales comprised of two opposite adjectives pairs, such as white-black, hot-cold, fast-slow, and so on. The two major aspects are valence (how good or negative a feeling is) and arousal (how intense an emotion is, ranging from drowsiness to enthusiasm) [Grekow, 2018].

Unlike dimensional emotion theory, the categorical or basic emotional theory is founded on a distinct collection of basic emotions and, in certain circumstances, secondary emotions derived from the primary ones. Ekman performed the most well-known and largely recognized study on fundamental emotions. According to Ekman [Ekman, 1992], there are six fundamental emotions that may be defined: fear, joy, disgust, surprise, anger, and sad, as shown in Figure 1(a). Plutcnik proposes yet another categorical emotion paradigm. Plutcnik decided to display his model using eight fundamental emotions via a wheel resembling the famous color wheel, as seen in Figure 1. (b) [Seyeditabari, Tabari and Zadrozny, 2018].

On this wheel, Plutchik's paradigm presents the eight fundamental emotions in opposite pairs (anticipation versus surprise, anger versus fear, pleasure vs. sadness, trust versus disgust). The distance between each emotion's location and the middle of the wheel represents the activation of the associated emotion [Kołakowska *et al.*,2015]. According to the appraisal emotion theory, emotions emerge from one's observations and cognitive assessments of their surroundings. Individual differences in emotional responses to the same situation are accounted for by the theory. Its use to automatic emotion identification, on the other hand, is still in its early phases [So *et al*., 2015].



(A) **(B)**

Figure 1 Emotion theory, (A) Ekman model, (B) Plutchik emotion model

Information Channels of Emotion Recognition

The interactive communications can be classified into non-verbal and verbal [Alonso-Martín *et al*., 2013]. Verbal information consists of spoken words, whereas non-verbal information consists of information other than spoken words, such as face images, speech, and body language. Non-verbal behaviors, according to Ekman [Ekman, 1992], are the most important means of transmitting feelings. Mehrabian backs up this claim. [Mehrabian, 1971] who discovered that non-verbal correspondence is the most common type, with manner of speaking and body language accounting for 38 % and 55 % of emotional information, in both, while the remaining 7% is made up of spoken words (the "38 %– 55 %–7 % rule"). In this way, the examination of non-verbal information channels plays a critical role in the comprehension and synthesis of emotions. [Alonso-Martín *et al*., 2013].

Recently, researchers have focused on the use of non-verbal communication, such as eye movements, facial expressions, and physiological reactions, which use a singular channel of information, known as "mode" for example, speech, facial expression, and body language. Approaches that only use speech or visual data separately frequently fail in actual situations. Shadow, Occlusions, illumination conditions, and a variety of other factors, for example, are common conditions in which the precision of systems for visual frameworks decreases. Similarly, in speech frameworks, environmental noise or maybe a person traveling while speaking are regarded as error sound sources. Emotions can also be conveyed through more than one mode at same time to solve complications of a one modality. In multimodal emotion-based communication, the person can communicate his emotional state using a variety of information channels [Martinez-Martin and del Pobil, 2017].

In contrast to a single mode in which each channel displays data in an emotion recognition system, the majority of multimodal frameworks use certain channels as redundancy data. This redundant data is useful in real interactions, where mistakes caused by noisy sound or a partially occluded face, for example, can be reduced [Cid, Manso and Núñez, 2015].

As a result, many authors have focused on multi-modal approaches that locate user emotion from multiple data sources, such as face motion, speech, text and many others [Cid, Manso and Núñez, 2015]. The emotion recognition systems are divided into two groups in this review paper, according to the number of data channels, there are two types of systems: (1) single channel (uni-model) and (2) multi-channels (multi-modal).

Single Channel to Emotion Recognition

Many related works did not merge the information channels into a single model to recognize of human emotion; instead, various data sources (facial expressions, speech, text etc.) are usually used independently of one another. The subsections that follow explain various information channels that can be used to identify human emotions [Gunes and Pantic, 2010]. We will focus here on only two information channels: the facial expression and speech.

Speech

Human speech is focused with computerized human audio and speech processing topics because it communicates many types of information among humans. Computerized voice processing is dependent on a diverse set of disciplines and methodologies. A broad understanding of these subjects is required for effective advancement in its application domains such as speech recognition, language processing for speech comprehension, analysis, phonetics, and so on [Hatem et al., 2021]. The most convenient, easiest, and natural method to convey emotions is through speech. Human speech involves both content and attitude information. The Speech emotion recognition is a component of effective computing, detects emotional states through talk and reveals the attitudes suggested in the spoken language. Among the most challenging aspects of emotion recognition through speech is differentiating emotions across various languages, especially in the situation of mixed languages. Individuals may employ mixed terms from many languages, such as English words with words from another language (culture related) [Deriche and Abo absa, 2017].

Despite the fact that several speech features have been examined in the identification of speech emotions, the researchers have yet to determine the optimum speech characteristics for this purpose. This is because the retrieved features for the various emotions are similar. However, there seem to be major groups of speech features to consider of which including Continuous features that are significant in providing speakers' emotional indicators and hence widely used during speech emotion recognition. Most researchers think that feature of effective continuous like energy and pitch accurately represent the emotional content of an utterance. The speaker's arousal level, for example, influences the overall energy level as well as the length and period of speech stops. Energy-related features, formants, pitch and timing-related features are some of the most well-known continuous acoustic features [Kunxia Wang *et al.*, 2015]. The pitch of a spoken signal is an essential feature. A speech's pitch indicates the lowness and highness of its tone. The pitch characteristics rose with high-arousal feelings like surprise and happiness, but reduced in low-arousal emotions like fear and sadness [Sondhi *et al*., 2015]. The auditory resonance of the human vocal tract is known as formant in phonetics. The amplitude peaks in the speech's frequency spectrum can be found and extracted. Timing-related characteristics reveal the distribution of duration-related factors like speech rate and the percentage of spoken compared unvoiced portions. With high-arousal emotions, timing features rise, while with low-arousal emotions, timing features decrease [Sondhi *et al*., 2015]. Zhu *et al.* (2017) [Zhu *et al*., 2017] retrieved zero crossing rate, format, pitch and shorter - term energy characteristics from a voice signal. For classification, the Deep Belief Network & SVM are then integrated. Experiments were carried out utilizing a database created by the Chinese academy of sciences, with a 95.8% accuracy rate. Spectral features are the second type of speech feature. It has been discovered that the emotional state of an utterance influences a distribution of spectral energy across the frequency range of the audio signals. For example, it has been discovered that a voice signal with a happy emotion has a high energy level in the higher frequency region, but a signal with a sad emotion has a low energy level at the same frequency range. However, to further exploit the spectral distribution throughout the audible frequency range, the calculated spectrum is frequently processed through with a banks of band-pass filters. The outcomes of these filters are then used to construct spectral features. Mel Frequency Cepstral Coefficients abbreviated with (MFCCs) are an example of spectral features of speech [Sondhi et al., 2015]. Chavhan *et al.* (2015) [Chavhan, Yelure and Tayade, 2015] proposed an automated Speech Emotion Recognition approach that retrieved MFCCs and Mel Energy Spectrum Dynamic Coefficients abbreviated with (MEDCs) from speech signals. For the classification job, the LIBSVM with Radial Basis Functions (RBF) kernel was utilized. The accuracy attained using the Berlin database was 93.75 %. Despite the fact that numerous research projects have been completed and several applications have been developed, audio emotion recognition stay a difficult task [Pervaiz and Ahmed, 2016]. The key reasons for this are the complexity in determining which speech characteristics are information-rich and what is not, the modifiability of expression for same emotion, probability of much more than one emotion in a same expression, and the differences in talking styles, sentences, rates and speakers and rates. As an example of speech emotion recognition problems, joyful and angry both share acoustic characteristics such as amplitude, pitch, and the number of times that speech passes the zero pivot. Similarly, sadness and fear have certain characteristics. As a result, difficulties arise when recognizing these two pairs of emotions [Pervaiz and Ahmed, 2016]. To overcome these challenges, the researchers sought to employ additional approaches in order to identify the best resolution to the acoustic feature extraction job. Additionally, identifying spoken language may aid in improving speech emotion detection identification. Deriche and Abo-absa (2017) [Deriche and Abo absa, 2017] proposed a two-stage speech emotion detection method that begins by trying to identify the language of the voice, and also a separate emotion identification system is built for every language to select the type of emotion (happy, neutral, sad and angry) from the dedicated language. The discriminated features were extracted using the Wavelets transform, and the Hidden Markov Model was then utilized to follow the changes in the wavelets-based feature set to detect the spoken language. Following language recognition, a collection of characteristics from speech, such as MFCCs and pitch, is retrieved, and a neural network is also used to predict the emotion class. The suggested system's total accuracy was 93%.

Furthermore, Tian and Watson (2018) [Tian and Watson, 2018] suggested an emotion identification system based on intrasegmental characteristics derived from lengthy monophthongs in continuous speech data. Initially, 36 vocal tract characteristics and 11 vowel source characteristics were retrieved, and an optimal selection was chosen using Maximum Relevance Minimal Redundancy Backward Wrapping which abbreviate (MRMRBW). The JL corpus was utilized to test the system's performance. Regardless of vowel type, a recognition accuracy of 70.5 % was attained for five different emotion. Table 1 demonstrate an overview of the works that use only of speech.

Table 1. Overview of the works that use only of speech

Authors	**Type of dataset**	**Methods**	**Accuracy**
[BANDELA and KUMAR, 2019]	EMO-DB and IEMOCAP	(semi-NMF) with singular value decomposition (SVD), and K-NN, SVM was used for classification.	89.3 for (KNN) 90.12 for (SVM)
[Wang *et al.*, 2020]	EESDB and EmoDB	wavelet packet coefficient (WPC), Sequential Floating Forward Search (SFFS) method, and used Linear SVM (LSVM) for classification	60.7 for (EESDB) and 79.5 for (EmoDB)
[Togootogtokh and Klasen,2021]	RAVDESS dataset	deep transfer learning	100%
[Zhu et al., 2017]	Using the database created by the Chinese Academy of Sciences	Format, MFCCs, pitch, energy and zero crossing rate features, SVM and DBN for classification	95.8%
[Chavhan, Yelure and Tayade, 2015]	Berlin database	LIBSVM with RBF kernel for classification, MFCCs and MEDC for extracting features	93.75%
[Mekruksavanich, Jitpattanakul and Hnoohom, 2020]	SAVEE, TESS, RAVDESS CREMA-D, and THAI	MFCC for feature extraction and CNN for classification	TESS: 55.71%, SAVEE: 65.83%, RAVDESS: 75.83%, CREMA-D: 65.77%, and THAI: 96.60%

i. Facial Expressions

Facial expressions are among the most important pieces of data in every face-to-face interaction. Thus, it is fair to say that the study of facial impressions has piqued the interest of many people for many decades, with numerous applications. Face emotional recognition is simply a pattern recognition issue that includes determining the best feature set from of the facial data being examined [Mehta, Siddiqui and Javaid, (2018)]. Various approaches are used to perform feature extraction tasks, which may be divided to two major groups: Appearance and Geometric features [Wu, Lin and Wei, 2014]. During any type of facial emotion, the appearance features show for a brief moment in the face (for example, bulges, the presence of specific facial wrinkles, forefront and the texture of the facial skin in regions surrounding the eyes and mouth). To identify the discriminative feature vector, transform filters like Haar wavelets and integral image filters are used to these regions [Ping Tian, 2013]. Liu et al. (2017) [Liu *et al.*, 2017] used Local Binary Pattern (LBP), 2D-Gabor, and a multiclass Extreme Learning Machine (ELM) classifier to detect the seven fundamental emotions in real-time facial expression recognition. On the Japanese Female Facial Expression (JAFFE) public dataset, the algorithm achieved an accuracy of 80%. But, appearance features are not resistant to diverse facial changes (such as scale of the face, appearing face region, and head region orientation, among others) and do not perform well with noise images.

As for the geometrical features, the position of important components of face like the eyebrows, mouth, eyes, and nose is monitored, and the relationship of geometric between specific key locations on the facial (e.g., angles, shapes, and distances) is considered. Murugappan and Mutawa (2021) [M. and A., 2021] suggested a triangulation technique for extracting a collection of geometric characteristics to identify six emotional expressions (anger, sadness, suriprise, fear, joy, and disgust) employing computer-generated markers. Haar-like characteristics are used to recognize the subject's face. In an automatic fashion, a mathematical model was applied to the locations of eight markers in a predetermined region also on subject's face. Five triangles are produced by rearranging the locations of eight markers like an border of each triangle, Later, the Lucas Kanade motion algorithm tracks these eight markers indefinitely. The circumference of an inscribed circle (ICC), area of the triangular (AoT), and an inscribed circular area of a triangles (ICAT) are retrieved as characteristics to identify face expressions. Using various forms of machine learning algorithms, these characteristics are utilized to discern six facial emotions. Using a Random Forest classifier, the greatest mean classification rate was 98.17 %. Geometric features are resistant to many facial changes such as face scaling, appearing face region, face position within the picture, and head area direction. Therefore, the extraction of these features might be deemed costly in terms of computation since it needs reliable and precise approaches for facial feature tracking and detection [Sumathi, Santhanam and Mahadevi, 2012].

In general, the geometric and appearances features methods have demonstrated the difficulties in extracting a feature vector that can reliably differentiate the emotion type from a facial image. As a result, the researchers sought to employ techniques other than the two basic kinds to extract discriminating features, first, use Ekman Action Units (AUs). AUs imitate the movement of face muscles during facial expressions. A combination of facial action units can be used to identify the emotion state [Ghayoumi and Bansal, 2016]. Hsu et al. (2017) [Hsu, Huang and Huang, 2017] identified AUs using the Gabor filter and the Support Vector Machine (SVM). Then, based on the detected AUs, a random forest classifier was employed to recognize the emotional state. Experiments using the Cohen-Kanade+ database revealed that the system can recognize facial emotions with a 95% accuracy. But, AUs recognition is a difficult task due to a variety of variables such as lighting changes, posture variations, and individual subject variances. Second, use the deep learning techniques. Deep learning algorithms have found widespread use in face emotion recognition tasks as neuro-based learning algorithms have advanced. Deep learning is the use of multi-layer multi-neuron neural networks to accomplish learning tasks such as clustering, classification, regression, decoding, encoding and among others [Hatcher and Yu, 2018]. Deep learning approaches are divided into two architectures: Convolution Neural Network (CNN) and Recurrent Neural Network (RNN). CNN is a fed forward network that is mostly utilized in image processing. The number of hidden layers employed between the input and output layers determines the strength of the CNN. Each layer extracts a collection of features. A sequence of filters are applied to the input image to produce feature maps. Each filter traverses the whole image, multiplying its weights by the input values. RNN employs the back propagation principle, which means that it examines previous inputs in addition to the current inputs. With the assistance of a memory cell, RNNs can handle sequential data. RNNs are built on the premise that people do not conceive from of the scratch every time [Prabha and Umarani Srikanth, 2019]. Z. Zhang (2018) [Zhang, 2018] utilized CNN to solve the challenge of emotion recognition from facial expressions. To solve difficult environments such as unclear face details and poor lighting, the author extracted numerous input features and utilized mask loss to focus also on correct local face features. The Recognition of Emotion in the Wild Challenge, Static Facial Expression Recognition sub-challenge (SFEW) database has been used as test material, and the current proposal outperformed the baseline results by approximately 35.38%. Deep learning approaches are more accurate and predictive, as well as more powerful and flexible [Hatcher and Yu, 2018]. Despite these advancements, choosing the best features and settings for deep learning parameters remains a difficult task [Sharma, 2019]. Table 2 demonstrate an overview of the works that use only of facial expressions.

Table 2 Overview of the works that use only of facial expressions

Authors	Type of dataset	Methods	Accuracy
[Mehendale, 2020]	CMU, NIST datasets and Caltech faces, Cohn–Kanade expression	Convolutional neural networks are used to recognize face emotions (FERC)	96%
[Zhang, 2018]	SFEW dataset	Fusion of several input features with CNN to classify different emotions	Over the baseline values, there was a 35.38% improvement.
[Ahmed *et al.*, 2019]	N/A	convolutional neural network & data augmentation	96.24%
[Deriche and Abo absa, 2017]	N/A	To detect the language, use HMM. MFCCs are utilized in conjunction with NN to determine the emotion category.	Accuracy of about 93%
[Hsu, Huang and Huang, 2017]	CK+ database	The Gabor filter & SVM are used to identify AUs, while the randomized forest classifier is used to distinguish the emotional state.	95%
[Poulose *et al.*, 2021]	Collected user sample emotions from 9 people.	deep learning model (Xception architecture)	3.33% accuracy improvement than the conventional approach

Multi-channel to Emotion Recognition

Multimodal or (multiple information channels) emotion recognition gathers data from the more than one source. In general, three forms of data fusion techniques are: (a) data-level fusion technique, (b) decision-level fusion technique and (c) feature-level fusion technique. Data-level fusion could be accomplished by combining several physical signals of comparable type for example (two videos of two cameras, two EEG signals, etc.). Because data-fusion necessitates that various modalities have always had the same signal and nature characteristics. This fusion is not practical [Rabie, 2010]. Using feature-level fusion, a single classifier takes all characteristics derived from each source as input and uses decision algorithms to make a conclusion about the user's emotion. The feature-level fusion approach takes use of the correlation of selected features in multiple modalities. However, it has been criticized for ignoring the differences in temporal organization, size, and measurements across various characteristics. It also necessitates synchronization among modalities. It is also more complex and costly than combining at the decision level [Alonso-Martín et al., 2013]. A classifier is allocated to each channel in the decision-level fusion technique to make a judgment about the type of identified emotion. A decision rule must be provided in this function to determine the final emotion classification. The benefits of decision level fusion are simple, no need for synchronizing among modes, and it does not necessitate large computational requirements. For all these factors, decision-level fusion methods are commonly employed by researches in the area of multimodal emotion detection [Alonso-Martín et al., 2013]. Figure 2 below illustrates the three main fusion methods mentioned above.

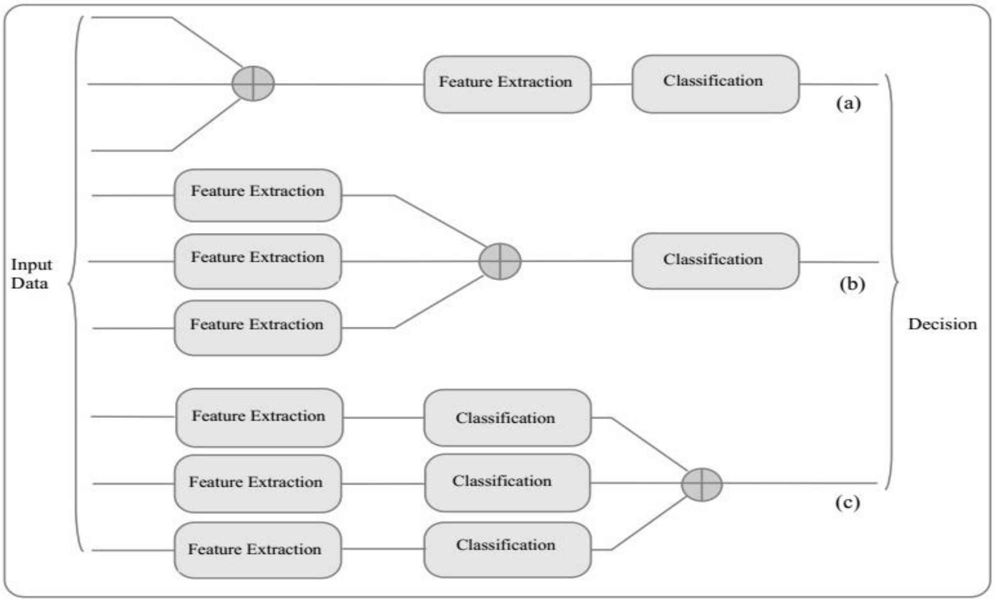


Figure 2 Three main fusion methods. In (a) data-level fusion technique, (b) feature-level fusion technique and (c) decision-level fusion technique

As seen in table 3, summary of some of the multi-source data (multimodal) combinations mentioned in the literature that include face and speech modes, text and speech modes, speech, face, and gesture modes, Bio-Signal and so on.

Table 3 Summary of some of the multi-source data (multimodal) of emotion recognition studies

Author	**Type of dataset**	**Modalities**	**Methods**	**Accuracy**
[Mittal et al., 2020]	IEMOCAP and CMU-MOSEI	Face+ Speech + Text	Deep learning model	82.7% for IEMOCAP and 89.0% for CMU-MOSEI
[Pérez *et al.*, 2018]	a collection of video and audio from a number of variety of users	Face + Speech	Corners and borders features and use CNN for classification, features of audio are MFCCs	86.4%.
[Krishna and Patil, 2020]	IEMOCAP	Speech +Text	Raw waveform and cross-modal attention based convolutional neural networks	72.82%
[Nguyen *et al*., 2018]	FABO and ENTERFACE database	Face + speech	3D convolution neural networks and deep belief networks	92.24% for FABO and 90.85% for ENTERFACE
[Cid, Manso and Núñez, 2015]	SAVEE database	Speech + Face	The dynamic and edge features, speech descriptors and Bayesian classifier	N/A
[Tan *et al.*, 2021]	Fer2013 and SEED-IV dataset	Face + EEG	The facial expressions were classified using CNN, the characteristics of the EEG signals were classified using SVM, and the multimodal recognition results were calculated using the Monte Carlo technique.	83.33%

Discussions

Despite the fact that several techniques to developing emotion recognition systems have been presented, the resulting systems still have significant limitations. As a result, among some of the recommendations for future productions are that the majority of emotion recognition systems are designed to classify the above-mentioned emotions in some literatures using the categorical emotion concept. However, some applications that require precision and sensitivity must rely on arousal or dimensional theories. As a result, more efforts are required to develop an emotion detection system based on such emotion theories. In the literatures, only basic emotions have been discussed. However, in real life, people might exhibit many emotions (mixed emotion) at the same moment. For example, a person may exhibit both surprise and happiness. As a result, when designing solutions that address this issue, this must be considered. The deep learning approaches has opened up a new study avenue. These approaches may be used by researchers to produce more promising studies in the area of emotional recognition. Some new techniques for emotion recognition have been suggested. However, these techniques may take a very long time to perform, making them inefficient for usage in real time implementations. Most research did not take into account dynamic environment circumstances, for example, most researches training their systems on a weak datasets, which leads low flexibility with changing environments. The created systems must be evaluated in more realistic situations, such as more loud surroundings, with more participants of various cultures, ages, and backgrounds, and so on. With the rapid development of sensor technologies that can record various bio-data signals, it has recently been possible to construct a multimodal emotional system by combining several physiological signals. This might lead to a significant advance in emotion detection systems.

Conclusion

This study provides a review of emotion recognition systems. We also looked at various information channels of emotion recognition and the latest developments in the design of various emotion recognition systems for each one or multiple modality. Finally, recommendations in this field have been suggested as well as trends.

References

Ahmed, T.U., Hossain, S., Hossain, M.S., ul Islam, R., & Andersson, K. (2019). Facial expression recognition using convolutional neural network with data augmentation. *In Joint 8th International Conference on Informatics, Electronics & Vision (ICIEV) and*

3rd International Conference on Imaging, Vision & Pattern Recognition (icIVPR), 336-341. http://doi.org/10.1109/ICIEV.2019.8858529

Alonso-Martín, F., Malfaz, M., Sequeira, J., Gorostiza, J. and Salichs, M. (2013). A Multimodal Emotion Detection System during Human–Robot Interaction. *Sensors, 13*(11), 15549–15581.

Bandela, S.R., & Kumar, T.K. (2019). Speech emotion recognition using semi-NMF feature optimization. *Turkish Journal of Electrical Engineering & Computer Sciences, 27*(5), 3741–3757.

Chavhan, Y.D., Yelure, B.S., & Tayade, K.N. (2015). Speech emotion recognition using RBF kernel of LIBSVM. *In 2nd international conference on electronics and communication systems (ICECS),* 1132-1135.

Cid, F., Manso, L.J., & Núnez, P. (2015). A novel multimodal emotion recognition approach for affective human robot interaction. *Proceedings of fine,* 1-9.

Deriche, M. (2017). A two-stage hierarchical bilingual emotion recognition system using a hidden Markov model and neural networks. *Arabian Journal for Science and Engineering, 42*(12), 5231-5249.

Ekman, P. (1992). An argument for basic emotions. *Cognition & emotion, 6*(3-4), 169-200.

Gharavian, D., Bejani, M., & Sheikhan, M. (2016). Audio-visual emotion recognition using FCBF feature selection method and particle swarm optimization for fuzzy ARTMAP neural networks. *Multimedia Tools and Applications, 76*(2), 2331–2352.

Ghayoumi, M., & Bansal, A.K. (2016). Unifying geometric features and facial action units for improved performance of facial expression analysis. arXiv preprint arXiv:1606.00822. 259-266.

Grekow, J. (2018). From Content-based Music Emotion Recognition to Emotion Maps of Musical Pieces. Studies in Computational Intelligence. Cham: Springer International Publishing.

Gunes, H., & Pantic, M. (2010). Automatic, Dimensional and Continuous Emotion Recognition. *International Journal of Synthetic Emotions, 1*(1), 68–99.

Hatcher, W.G., & Yu, W. (2018). A Survey of Deep Learning: Platforms, Applications and Emerging Research Trends. *IEEE Access, 6,* 24411–24432. http://dx.doi.org/10.1109/ACCESS.2018.2830661.

Hatem, A.S., Adulredhi, M.J., Abdulrahman, A.M., & Fadhel, M.A. (2020). Human Speaker Recognition Based Database Method. *In International Conference on Intelligent Systems Design and Applications,* 1145-1154.

Hsu, S.C., Huang, H.H., & Huang, C.L. (2017). Facial expression recognition for human-robot interaction. *In First IEEE International Conference on Robotic Computing (IRC),* 1-7. IEEE. http://dx.doi.org/10.1109/IRC.2017.12

Kołakowska, A., Landowska, A., Szwoch, M., Szwoch, W., & Wróbel, M.R. (2015). Modeling emotions for affect-aware applications. *Information Systems Development and Applications,* 55-69.

Krishna, D.N., & Patil, A. (2020). Multimodal Emotion Recognition Using Cross-Modal Attention and 1D Convolutional Neural Networks. *In Interspeech,* 4243-4247.

Wang, K., An, N., Li, B. N., Zhang, Y., & Li, L. (2015). Speech emotion recognition using Fourier parameters. *IEEE Transactions on affective computing, 6*(1), 69-75.

Liu, Z., Wu, M., Cao, W., Chen, L., Xu, J., Zhang, R., Zhou, M., & Mao, J. (2017). A facial expression emotion recognition based human-robot interaction system. *IEEE/CAA Journal of Automatica Sinica, 4*(4), 668–676. http://dx.doi.org/10.1109/JAS.2017.7510 622.

Murugappan, M., & Mutawa, A. (2021). Facial geometric feature extraction based emotional expression classification using machine learning algorithms. *Plos one, 16*(2), e0247131. https://doi.org/10.1371/journal.pone.0247131.

Martinez-Martin, E., & Del Pobil, A.P. (2017). Object detection and recognition for assistive robots: Experimentation and implementation. *IEEE Robotics & Automation Magazine, 24*(3), 123-138.

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (FERC). *SN Applied Sciences, 2*(3), 1-8. https://doi.org/10.1007/s42452-020-2234-1

Mehrabian, A. (1971). *Silent messages. Open WorldCat.* Belmont, California: Wadsworth Publishing Company.

Mehta, D., Siddiqui, M., & Javaid, A.Y. (2018). Facial Emotion Recognition: A Survey and Real-World User Experiences in Mixed Reality. *Sensors, 18*(2). https://doi.org/10.3390/s18020416

Mekruksavanich, S., Jitpattanakul, A., & Hnoohom, N. (2020). Negative Emotion Recognition using Deep Learning for Thai Language. *In Joint International Conference on Digital Arts, Media and Technology with ECTI Northern Section Conference on Electrical, Electronics, Computer and Telecommunications Engineering (ECTI DAMT & NCON),* 71-74.

Mittal, T., Bhattacharya, U., Chandra, R., Bera, A., & Manocha, D. (2020). M3er: Multiplicative multimodal emotion recognition using facial, textual, and speech cues. *In Proceedings of the AAAI Conference on Artificial Intelligence, 34*(2), 1359-1367.

Nguyen, D., Nguyen, K., Sridharan, S., Dean, D., & Fookes, C. (2018). Deep spatio-temporal feature fusion with compact bilinear pooling for multimodal emotion recognition. *Computer Vision and Image Understanding, 174,* 33-42. https://doi.org/10.1016/j.cviu.2018.06.005.

Pérez, A.K., Quintero, C.A., Rodríguez, S., Rojas, E., Peña, O., & De La Rosa, F. (2018). Identification of Multimodal Signals for Emotion Recognition in the Context of Human-Robot Interaction. *Intelligent Computing Systems,* 67–80. http://dx.doi.org/10.1007/978-3-319- 76261-6_6.

Pervaiz, M., & Khan, T.A. (2016). Emotion recognition from speech using prosodic and linguistic features. *International Journal of Advanced Computer Science and Applications, 7*(8), 84-90.

Ping Tian, D. (2013). A Review on Image Feature Extraction and Representation Techniques. *International Journal of Multimedia and Ubiquitous Engineering, 8*(4), 385-396.

Poulose, A., Reddy, C.S., Kim, J.H., & Han, D.S. (2021). Foreground Extraction Based Facial Emotion Recognition Using Deep Learning Xception Model. *In Twelfth International Conference on Ubiquitous and Future Networks (ICUFN),* 356-360. http://doi.org/10.1109/ICUFN49451.2021.9528706

Prabha, M.I., & Srikanth, G.U. (2019). Survey of sentiment analysis using deep learning techniques. *In 1st International Conference on Innovations in Information and Communication Technology (ICIICT)*, 1-9. http://dx.doi.org/10.1109/ICIICT1.2019.8741438.

Rabie, A. (2010). *Audio-Visual Emotion Recognition for Natural Human-Robot Interaction*. Ph.D. Dissertation, Bielefeld University.

Seyeditabari, A., Tabari, N., & Zadrozny, W. (2018). *Emotion Detection in Text: a Review*. arXiv:1806.00674 [cs]. https://arxiv.org/abs/1806.00674

Sharma, O. (2019). Deep Challenges Associated with Deep Learning. 2019 International Conference on Machine Learning, Big Data, Cloud and Parallel Computing (COMITCon), 72- 75.

Sharma, O. (2019). Deep challenges associated with deep learning. *In international conference on machine learning, big data, cloud and parallel computing (COMITCon)*, 72-75. http://dx.doi.org/10.1109/COMITCon.2019.8862453

So, J., Achar, C., Han, D., Agrawal, N., Duhachek, A., & Maheswaran, D. (2015). The psychology of appraisal: Specific emotions and decision-making. *Journal of Consumer Psychology, 25*(3), 359-371.

Sondhi, S., Khan, M., Vijay, R., Salhan, A.K., & Chouhan, S. (2015). Acoustic analysis of speech under stress. *International journal of bioinformatics research and applications, 11*(5), 417-432.

Sumathi, C.P., Santhanam, T., & Mahadevi, M. (2012). Automatic Facial Expression Analysis: A Survey. *International Journal of Computer Science and Engineering Survey, 3*(6), 47-59.

Tan, Y., Sun, Z., Duan, F., Solé-Casals, J., & Caiafa, C.F. (2021). A multimodal emotion recognition method based on facial expressions and electroencephalography. *Biomedical Signal Processing and Control, 70*. https://doi.org/10.1016/j.bspc.2021.103029

Tian, L., & Watson, C.I. (2018). Emotion Recognition Using Intrasegmental Features of Continuous Speech. *In Proc. of the 17th International Australasian Conference on Speech Science and Technology*, 61-64.

Togootogtokh, E., & Klasen, C. (2021). DeepEMO: Deep Learning for Speech Emotion Recognition. arXiv preprint arXiv:2109.04081.

Wang, K., Su, G., Liu, L., & Wang, S. (2020). Wavelet packet analysis for speaker-independent emotion recognition. *Neurocomputing, 398,* 257-264.

Wu, C.H., Lin, J.C., & Wei, W.L. (2014). Survey on audiovisual emotion recognition: databases, features, and data fusion strategies. *APSIPA transactions on signal and information processing, 3,* 1-18. http://doi.org/10.1017/ATSIP.2014.11

Zhang, Z. (2018). Deep Face Emotion Recognition. *In Journal of Physics: Conference Series, 1087*(6). http://dx.doi.org/10.1088/1742-6596/1087/6/062036

Zhao, S., Jia, G., Yang, J., Ding, G., & Keutzer, K. (2021). Emotion Recognition From Multiple Modalities: Fundamentals and methodologies. *IEEE Signal Processing Magazine, 38*(6), 59-73. https://arxiv.org/abs/2108.10152

Zhu, L., Chen, L., Zhao, D., Zhou, J., & Zhang, W. (2017). Emotion recognition from Chinese speech for smart affective services using a combination of SVM and DBN. *Sensors, 17*(7).