Evaluating Methods for Emotion Recognition based on Facial and Vocal Features

Matthias Ley and Maria Egger
Center for Health and Bioresources
AIT Austrian Institute of Technology GmbH
Wiener Neustadt
Austria

Sten Hanke
Institute of eHealth
University of Applied Sciences - FH JOANNEUM GmbH
Graz
Austria

Abstract

An automatized emotion recognition based on Information and Communication Technology (ICT) is currently of high interest and opens possibilities for various applications. Emotions are a key feature in communication and the possibility of recognizing emotions may advance Human Computer Interaction. Furthermore unobtrusive emotion recognition can be used to determine levels of stress while handling machines and vehicles and helps to reduce the risk of car crashes or accidents at work. The e-health sector benefits from emotion recognition to monitor mental states of patients at home and to improve recovery rates through a more personalized care. Emotion recognition might be useful to predict emotions from patients who are unable to express their emotions, for example patients suffering from autism, Parkinsons or locked-in syndrome. At the moment several methods for an ICT based emotion recognition exist. The methods vary in terms of precision, usability, application area as well as number of emotions which can be detected. An overview of methods used for emotion recognition and the current state of the art is given in this paper. The paper is focusing on evaluating existing tools for emotion recognition based on facial features as well as vocal features in voice interactions. The methods have been tested for its usability in the field of e-health and applications for elderly care. A case study was conducted to elicit emotions in 20 participants after presenting them an affective video based on rated picture and sound databases. Furthermore, video recordings of the frontal face as well as spoken texts were used for testing the usability of the three projects. The results suggest that facial recognition works best for the emotion happiness while voice recognition works best for the emotion anger.

In: E. Calvanese Strinati, D. Charitos, I. Chatzigiannakis, P. Ciampolini, F. Cuomo, P. Di Lorenzo, D. Gavalas, S. Hanke, A. Komninos, G. Mylonas (eds.): Proceeding of the Poster and Workshop Sessions of AmI-2019, the 2019 European Conference on Ambient Intelligence, Rome, Italy, 13-11-2019, published at http://ceur-ws.org

1 Introduction

The ability to recognize emotions is one of the criteria for emotional intelligence and therefore an important part of human intelligence. Machine intelligence needs to include emotional intelligence to recognize human affective states. Emotion recognition is therefore fundamental towards advanced Human Computer Interaction. Recent research emphasizes on recognition of emotional reactions from non-verbal cues such as facial expressions, voice, gestures and bio signals [1] [3] [2]. Ekman and Friesen defined six basic emotions, which are valid for all ages and cultural differences [4]. In 1873 Wilhelm Wundt designed a novel three-dimensional emotion classification system describing the valence, arousal and intensity of emotions on three axes [5]. Most common is a two dimensional model derived from the circumplex model, where valence describes if the emotion is positive or negative and arousal if the emotion is more passive or active. Happiness, for example, is labeled as an emotion with both high valence and arousal [6]. A similar model called the Geneva Emotion Wheel is commonly used in recent studies on emotion recognition [7]. The methods mentioned for emotion recognition vary in terms of accuracy and number of emotions which can be detected. The application where the emotion recognition will be used needs to be considered. For Activities of Daily Living (ADL) emotion recognition based on wearable measurement systems (e.g. smart watches) to measure physiological bio-signals might be appropriate. For applications with a strong focus on video communication (e.g. in voice-video communication or interactions with an avatar or a robot equipped with a camera) facial and voice recognition is preferable. For a detailed analysis of emotion recognition methods respective to the application area see [18]. This paper is analysing methods for emotion recognition based on facial and vocal features. Three toolboxes and SDKs are tested in a simple setup to evaluate their usefulness for emotion recognition applications. The following technologies were tested: the commercial FaceReader [1] Software by Noldus Information Technology B.V., the Affdex SDK [2] by Affectiva as well as OpenVokaturi [3] by Vokaturi B.V.

1.1 State of the Art in Emotion Recognition

1.1.1 Overview Emotion Recognition

Emotion recognition can be achieved with various techniques. The right method depends on the application area as well as the to be analysed emotions. The accuracy for one certain emotion and method might not be reproducible with another method. Multi modal systems are used to compensate low accuracy rates for certain emotions. Furthermore, emotion recognition from bio signals (e.g. from electrodermal activity, respiration, skin temperature or electromyography) generally requires a multi modal setup. This is due to the fact, that most bio signals are not sufficient to describe emotions on a two dimensional scale in a uni modal setup. Electrodermal activity, for example, only allows the expression of the excitement level of a person (arousal). Table 1 shows an overview of possible accuracy, benefits and limitations of common methods for emotion recognition. Regarding the state of the art in emotion recognition, several review papers give more detailed insight into measurement setups and their classification accuracy [18] [19] [20] [21].

1.1.2 Emotion Recognition based on Facial and Vocal Features

Several studies use a multi-cue approach when building tools for emotion recognition. Data from facial expressions as well as from speech is merged to increase the classification rate [28], [29], [30]. For validation of the methods some studies validate their data against lab trials where participants are asked to pose certain emotions. It is questionable if validating against actors is a viable method because the process of emotion expression is complex and might not be reproducible when acted out by humans. This fact might be easier for facial and vocal cues and more inaccurate when concerning bio signals caused by the autonomous nervous system. More promising is a validation against labeled data such as rated video and sound databases, which is the preferred method for a comparable validation. For example Shan et al. [31] is validating several methods using local binary patterns in pictures against rated databases (Japanese Female Facial Expression (JAFFE) Database and the Cohn-Kanade database) [32]. Kahou et al. [33] uses a deep learning approach to classify 7 emotions from short clips of Hollywood movies based on visual face features. Tivatansakul et al. [34] uses emotion recognition based on facial expression to provide better health services. Furthermore there are emotion recognition approaches based on vocal features as well as from the context of the message (vocabulary, syntax and usage of words). For vocal

[1] https://www.noldus.com
[2] https://www.affectiva.com/product/emotion-sdk/
[3] https://vokaturi.com/

features, Zhang et al. [35] included the Mel frequency cepstrum coefficient, pitch and formants in his analysis to recognize six emotions.

Table 1: Overview of Common Modalities, Their Benefits, Limitations and Expectable Accurancy. Modalities: Electroencephalography, Facial Recognition, Speech Recognition, Electrocardiography, Electrodermal Activity, Respiration, Skin Temperature, Electromyography.

Modality	Benefits	Limitation	Accuracy
EEG	measurement of impaired patients	complex installation, high maintenance, lab conditions	58 % (4 emot.) [25]
FR	contact-less, multi person tracking	req. frontal face, camera, does not allow free movement	48 % (8 emot.) [1]
SR	contact-less, unobtrusive, casual	req. communication, prone to background noise	62 % (4 emot.) [2]
ECG	measurement with smart watch possible	prone to movement artifacts	78 % (6 emot.) [26]
EDA	good indicator for stress, distinction conflict and no conflict [27]	only arousal, influenced by temperature and sweat	67 % (2 states) [13]
RSP	simple setup, indication of panic, fear, depression [27]	difficult distinction for broader emotional spectrum, only used in multi modal setup	79 % (4 emot.) (EMG, ECG, RSP, EDA) [27]
SKT	versatile data acquisition (IR, through video, sensors)	only arousal, slow reacting bio signal, influenced by external temp.	significant difference betw. happy and sad [22]
EMG	measurement of impaired patients	only valence, difficult setup, lab conditions	linear effect for pos. and neg. affect [23] [24]

1.2 Research Question

Table 1 shows preferable methods depending on the use case and application area. The upswing of smart wearable devices and the vast interest in human-computer-interaction made contact-less measurement methods popular. This makes testing and verification of the wide variety of existing systems and projects for emotion recognition necessary. This paper should give an overview of the applicability of facial and voice recognition on the basis of different elicitation material (self recorded video data from the frontal face, existing video and sound data from rated/unrated sources). Ideally this information is useful for improving the human-computer-interaction in elderly care and e-health applications.

2 Methods

The following three projects were chosen for testing the usability of facial and vocal emotion recognition: To recognise emotions from facial features, the FaceReader by Noldus Information Technology B.V., Wageningen,

Netherlands as well as the Affdex SDK by Affectiva, Boston, Massachusetts were used. To analyse spoken texts taken from audio recordings, the OpenVokaturi SDK, Vokaturi B.V., Amsterdam, Netherlands was used.

2.1 Facial Recognition with FaceReader

Four basic emotions were chosen for investigation - contentment, disgust, sadness and happiness. A study was conducted at the AIT Austrian Institute of Technology GmbH, Wiener Neustadt, Austria as well as the University of Applied Sciences Technikum Wien, Vienna, Austria. A total of 23 data sets were recorded, based on 21 participants. The age of the participants ranged from 23 to 39 years. Males, females, people with glasses, beards and different skin colors were involved. One participant took part in an extended study to prove the reproducibility of the system. A video was shown to the participants based on rated picture (NAPS) and sound (DEAM) databases to elicit the desired emotions [8] [9]. The video consisted of four emotional phases with two minutes for each phase. Between these phases a relaxation phase ensures that the emotional effects and the reaction of the autonomous nervous system calmed down. The frontal face of the participants was recorded during the whole measurement protocol. The recorded videos of the participants faces were then used for the analysis with the FaceReader. The participants rated their emotional experience with a self assessment test (SAM) [10] at the beginning of each relaxation phase. The test involved rating of the participants arousal (from mild to intense) and valence (from unpleasant to pleasant) on a one to five scale. MATLAB by The MathWorks Inc., Massachusetts, United States of America, was used for the analysis and evaluation of the recorded data. The self assessed scores were correlated against the classification results from the FaceReader.

2.2 Facial Recognition with Affdex

To test the usability of the Affdex SDK for facial emotion recognition video material from the platform YouTube was used. Table 2 describes the videos used for the analysis. Contrary to the previous method this part of the research was based on publicly available videos of people showing emotions for various reasons (mostly acting) instead of having people react to emotional elicitation material. The videos were not chosen based on rated databases and were not annotated by experts. It was ensured that videos showing the frontal face were used and proper lighting conditions were met. The length of the videos ranged from short videos (4 to 11 seconds) to longer videos (1 to 3 minutes). One video file (ID3, table 5) was separated into eight smaller parts to only analyse relevant emotions instead of the original 30 emotions portrayed by the same actor (Breeze Woodson). The videos were rated individually and the average classification accuracy of seven emotions was calculated. The analysed emotions were contentment, surprise, anger, sadness, disgust, fear and joy. The Affdex SDK was implemented in a C# based project which was able to analyse the recorded video files as well as to analyse faces with a webcam in real time. The results were then exported and analysed in MATLAB.

2.3 Voice Recognition with OpenVokaturi

Similar to section 2.2 this part of the research was not based on self made recordings but on an annotated dataset. For testing the OpenVokaturi SDK an audio visual database called RAVDESS (Ryerson Audio-Visual Database of Emotional Speech and Song) was used [11]. The database contains recordings from 24 actors. For speech records, two statements were spoken in two intensities (normal and strong) in two repetitions. A Python script was written to analyse the sound files. The files had to be corrected in Audacity recording and editing software [12]. The files were imported and saved as a 16-bit mono WAV files without meta descriptions. After the classification of the files with the OpenVokaturi SDK was successful, the results were exported and further analysed in MATLAB.

3 Results

3.1 FaceReader

Table 3 shows the self rated arousal and valence scores from the participants after watching the emotional phases from the elicitation video. This result gives insight whether the material used was appropriate to elicit the desired emotions. Table 4 shows the average classification accuracy for one subject over three measurements (reproducibility measurement). The participant was instructed to force facial expressions based on the experienced emotion. Figure 1 shows the average classification accuracy for all 23 datasets. The accuracy is calculated over the length of two minutes for each emotional phase. The participants were instructed to behave naturally.

Table 2: YouTube videos for the analysis with the Affdex SDK. Displayed are the Video ID, Description of the Recording, Displayed Emotion, Length of the Video File and the Source of the YouTube Video.

ID	Description	Emotion	Length [m:s]	Source
1	Can You Watch This Without Smiling	Happy	01:11	[14]
2	Acting - Sad Scene	Sad	02:45	[15]
3	100 People Show Us What It Looks Like When They Cry	Sad	02:22	[16]
	30 EMOTIONS - Breeze Woodson			[17]
4	Excerpt from ID3	Angry	00:09	
5	Excerpt from ID3	Disgust	00:04	
6	Excerpt from ID3	Happy	00:07	
7	Excerpt from ID3	Laughing	00:11	
8	Excerpt from ID3	Pain / Ouch	00:04	
9	Excerpt from ID3	Pouty	00:04	
10	Excerpt from ID3	Sad	00:10	

Table 3: Self Assessed Arousal and Valence Scores from 1 to 5 after Watching the Elicitation Video (Mean $\pm$ Standard Deviation) and the Desired Scores for the Four Analysed Emotions.

Emotion	Arousal	Valence	Desired Ar.	Desired V.
Contentment	1.6 ± 0.8	4.4 ± 0.5	≤ 3	> 3
Sad	2.3 ± 1.2	1.6 ± 0.5	≤ 3	≤ 3
Disgust	4 ± 1.2	1.2 ± 0.5	> 3	≤ 3
Happy	3.7 ± 1.0	4.8 ± 0.3	> 3	> 3

Table 4: Average Classification Accuracy with FaceReader for Three Measurements of the Same Participant While Acting Out Emotions Excessively. Data is Expressed in Prediction Rates from 0 to 100 % for Each Emotional Phase.

	Cont. [%]	Sad [%]	Disg. [%]	Happy [%]
Measur. 1	16.0	45.8	6.6	**70.7**
Measur. 2	16.8	46.2	6.9	**73.8**
Measur. 3	18.6	67.1	5.7	**86.4**
Mean $\pm$ std	17.2 ± 1.3	53.0 ± 12.2	6.4 ± 0.6	$\mathbf{76.9 \pm 8.3}$

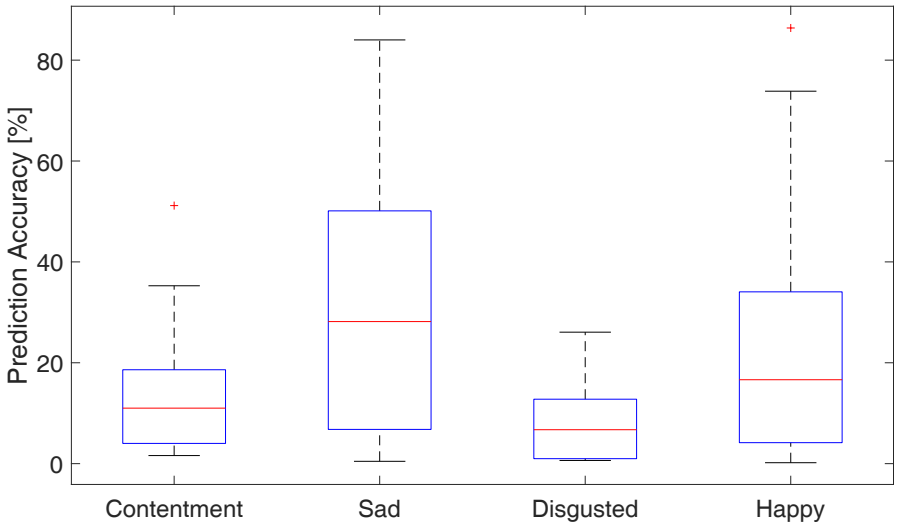

Figure 1: Average Classification Accuracy with FaceReader for All 23 Measurements While Acting Out Emotions Naturally. Data is Presented in Prediction Rates from 0 to 100 % for Each Emotional Phase.

3.2 Affdex

The following table 5 shows the results of the emotion recognition of YouTube videos with the Affdex SDK. The prediction results for contentment and fear were not displayed due to visibility and the fact, that their maximum accuracy (8.7 % for contentment, 1.7 % for fear) were exceptionally low.

Table 5: Analysis of the Video Files taken from YouTube. The Video ID, Displayed Emotion (Video), Detected Dominant Emotion and the Accuracy for Seven Emotions (Contentment, Surprise, Anger, Sadness, Disgust, Fear, Joy) in Percent from 0 to 100 are Displayed. Matching Dominant Emotions are Highlighted.

ID	Disp. Em.	Det. Em.	Sur.	Ang.	Sad	Disg.	Joy
1	**Happy**	**Joy**	9.8	1.7	0.5	5.0	**78.0**
2	Sad	Surprise	63.3	1.4	4.7	4.8	0.0
3	Sad	Anger	6.8	13.4	7.0	8.5	2.0
4	Angry	Sadness	0.0	21.3	45.7	3.1	0.0
5	Disgust	Anger	0.0	40.1	36.0	16.4	0.0
6	**Happy**	**Joy**	2.0	0.0	0.0	0.1	**81.3**
7	**Laughing**	**Joy**	3.3	5.6	3.8	31.0	**38.0**
8	**Pain**	**Anger**	0	**22.3**	4.6	4.0	0.0
9	Pouty	Disgust	0.2	0.0	0.0	0.5	0.0
10	Sad	Surprise	9.1	0.1	7.0	2.1	0.0
11	**Surprise**	**Surprise**	**37.8**	0	0.8	2.0	9.3

3.3 OpenVokaturi

The results of the emotion recognition through vocal features are presented in table 6. Displayed are the emotion labels of the audio recordings based on 192 samples for each emotion, the five emotional states (neutral, happy, sad, angry, fear) and the number of failed samples during the analysis. Similar to the other both methods, an overall prediction accuracy is expressed.

4 Conclusion

For the analysis with the FaceReader and the dataset gathered from one participant measured three times while excessively expressing the desired emotion, the best achieved classification accuracy occurred during the phase "Happy" with 76.9 % mean prediction accuracy (table 4). The classification accuracy for "Happy" is followed by "Sad" (53.0 %), "Contentment" (17.2 %) and eventually "Disgusted" (6.4 %). For the analysis of the whole 23

Table 6: Analysis of the RAVDESS sound files with the OpenVokaturi SDK. Displayed are five emotional phases (neutral, happy, sad, angry, fear) and the five prediction accuracy for the mentioned emotions as well as the number of failed samples during the analysis.

Emo.	**Neut.**	**Hap.**	**Sad**	**Ang.**	**Fear**	**Fail [n]**
Neut.	12.0	17.9	7.6	59.1	3.4	33
Hap.	2.2	41.8	0.8	48.5	6.8	6
Sad	5.3	31.7	9.2	35.7	18.1	49
Ang.	3.3	37.7	3.8	**42.0**	13.2	3
Fear	4.0	38.4	3.8	27.3	26.6	13

recordings (figure 1), the accuracy ranking for "Happy" and "Sad" are flipped compared to the previous analysis. The overall accuracy for all participants is lower than the accuracy for one subject forcing the emotions. This is probably based on the fact, that the participants were instructed to behave naturally. Furthermore, each picture taken from the rated database to create the video was presented for 15 seconds (eight pictures, total duration of two minutes per emotional phase). It seems natural, that the participants might express the emotion for a short amount of time and then return to a neutral facial state, which results in a lower overall accuracy. However, table 3 shows that the material used to elicit the emotions meets the requirements and that the participants experienced the desired emotions. In a further step, it might be beneficial to only analyse the dominant emotion over a certain time window, and to focus on the emotions happy and sad, as contentment and other emotions result in similar facial features to a neutral expression.

The usability of the Affdex SDK was based on the analysis of unrated video material and the sole decision of one person (author). It would be appropriate to refer to a rated database in future tests. For the video material used it could be observed that the recognition of the emotion happy works best (assuming that joy is closely related to happiness in terms of arousal and valence scores). Video ID 1 contains a compilation of people laughing and being happy in general. The high prediction rate of 78 % for ID 1 (table 5) suggests that the Affdex SDK works properly when detecting multiple faces (20 individual actors) (table 5. The video covers people from most ethnicities, eyeglass wearers, people with beards and includes males and females. Similar to the results with the FaceReader for three measurements on the same participant (table 4), the recognition of happy works better than for any other emotion. Besides happiness, one sample was correctly classified for surprise and pain (assumed to be anger) and no sad or disgust sample was correctly classified with the Affdex SDK. The unusual high recognition results for the emotion happy are most likely dependant on the unique facial characteristic for that emotion. In conclusion it can be said that emotion recognition with video material works best for the emotion happy across the two analysed systems.

After correcting the format of the audio files the script could not produce results for a number of files. The reason for this is unknown and might be caused by different recording methods of the database RAVDESS, the length or the quality of the audio files. Furthermore, the overall prediction accuracy is presented. Analysing the dominant emotion for each sample instead of the mean overall prediction accuracy for each emotion might give additional insight regarding which method is most applicable. Regarding table 6 it can be seen that even for the happy labeled sound samples the emotion could not be predicted correctly (angry dominates over happy detection). When listening to the audio files it seemed that some happy samples did not sound happy at all. This leads to the assumption that the RAVDESS database might have labeled their sound files misleadingly and that a different database might yield better results. The only positive classification was for the sound files labeled as angry (mean classification accuracy of 42.0 %). The results for the sound files labeled as sad were exceptionally bad (49 failed samples out of 192, mean classification accuracy of 9.2 % for sad). This might be caused by the fact, that sadly spoken sentences have less detectable features as when speaking in other emotional states. In general it seems that the classification for happy is confused with the classification for angry.

References

[1] L. Kessous, G. Castellano and G. Caridakis, *Multimodal emotion recognition in speech-based interaction using facial expression, body gesture and acoustic analysis*, J. Multimodal User Interfaces, vol. 3, no. 1, pp. 3348, 2010.

[2] K. Dai, H. J. Fell and J. MacAuslan, *Recognizing emotion in speech using neural networks*, in Proceedings of the 4th IASTED International Conference on Telehealth and Assistive Technologies, Telehealth/AT 2008, pp. 3136, 2008.

[3] C. Maaoui and A. Pruski, *Emotion recognition through physiological signals for human-machine communication*, Cut. Edge Robot., pp. 317333, 2010.

[4] P. Ekman and W. V. Friesen, *Constants across cultures in the face and emotion*, J. Pers. Soc. Psychol., vol. 17, no. 2, pp. 124129, 1971.

[5] W. Wundt, *Principles of physiological psychology*, 1873. East Norwalk, CT, US: Appleton-Century-Crofts., 1948.

[6] J. Posner, J. A. Russell, and B. S. Peterson, *The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology*, Dev. Psychopathol., vol. 17, no. 3, pp. 715734, Sep. 2005.

[7] V. Sacharin, K. Schlegel, and K. Scherer, *Geneva Emotion Wheel rating study (Report)*, Soc. Sci. Inf., 2005.

[8] A. Marchewka, . Zurawski, K. Jednorg, and A. Grabowska, *The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database*, Behav. Res. Methods, 2014.

[9] A. Marchewka, . Zurawski, K. Jednorg, and A. Grabowska, *The Nencki Affective Picture System (NAPS): Introduction to a novel, standardized, wide-range, high-quality, realistic picture database*, Behav. Res. Methods, 2014.

[10] J. D. Morris, *Observations: SAM: The self-assessment manikin: An efficient cross-cultural measurement of emotional response*, Journal of Advertising Research, vol. 35, no. 6. Advertising Research Foundation, US, pp. 6368, 1995.

[11] S.R. Livingstone, F.A. Russo, *The Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS): A dynamic, multimodal set of facial and vocal expressions in North American English*, PLoS ONE 13(5): e0196391, 2018.

[12] D. Mazzoni, *Audacity 2.3.2 recording and editing software*, copyright 1999-2019 Audacity Team, Web site: https://audacityteam.org/, free software distributed under the terms of the GNU General Public License, 2019.

[13] M. Ley, *Emotion recognition from facial recognition and bio signal analysis*, in fulfillment of the requirements for the degree of Master of Science in Engineering, University of Applied Sciences Technikum Wien, 2019.

[14] [BuzzFeedVideo], *Can You Watch This Without Smiling? [Video File]*, Retrieved from https://www.youtube.com/watch?v=f8OmSWxF6h8, 04.04.2016.

[15] [Artemis], *Acting - Sad Scene - Artemis [Video File]*, Retrieved from https://www.youtube.com/watch?v=9qRGBRYHO0g, 24.02.2014.

[16] [Cut], *Crying — 100 People Show Us What It Looks Like When They Cry — Keep it 100 — Cut [Video File]*, Retrieved from https://www.youtube.com/watch?v=vv2qnoUfjPU, 03.01.2017.

[17] [Breeze Woodson], *30 EMOTIONS — Breeze Woodson [Video File]*, Retrieved from https://www.youtube.com/watch?v=y2YUMPJATmg, 19.08.2016.

[18] M. Egger, M. Ley, and S. Hanke, *Emotion Recognition from Physiological Signal Analysis: A Review*, Electron. Notes Theor. Comput. Sci., 2019.

[19] J. Garcia-Garcia, V. Penichet, and M. Lozano, *Emotion detection: a technology review*, XVIII International Conference, 2017.

[20] S. Jerritta, M. Murugappan, R. Nagarajan, and K. Wan, *Physiological signals based human emotion Recognition: a review*, in 2011 IEEE 7th International Colloquium on Signal Processing and its Applications, pp. 410415, 2011.

[21] K. Wac and C. Tsiourti, *Ambulatory assessment of affect: Survey of sensor systems for monitoring of autonomic nervous systems activation in emotion*, IEEE Trans. Affect. Comput., 2014.

[22] L. Lundqvist, *Emotional responses to music: experience, expression, and physiology*, Psychology of Music, vol. 37, no. 1, p. 61-90, 2009.

[23] J. Larsen, *Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii*, Psychophysiology, vol. 40, no. 5, p. 776-785, 2003.

[24] R. Hazlett, *Measuring Emotional Valence during Interactive Experiences: Boys at Video Game Play*, CHI 2006 Proceedings, 2006.

[25] R. M. Mehmood and H. J. Lee, *A novel feature extraction method based on late positive potential for emotion recognition in human brain signal patterns*, Comput. Electr. Eng., 2016.

[26] F. Agrafioti, D. Hatzinakos, and A. K. Anderson, *ECG pattern analysis for emotion detection*, IEEE Trans. Affect. Comput., 2012.

[27] A. Haag, S. Goronzy, P. Schaich, and J. Williams, *Emotion Recognition Using Bio-sensors: First Steps towards an Automatic System*, Tutorial and research workshop on affective dialogue systems, p. 36-48, 2004.

[28] M. S. Hossain, G. Muhammad, M. F. Mohammed, B. Song and K. Al-Mutib, *Audio-Visual Emotion Recognition Using Big Data Towards 5G*, Mobile Networks and Applications, vol. 5, Springer New York LLC, p. 753–763, 2016.

[29] J. Yan, W. Zheng, Z. Cui, C. Tang, T. Zhang and Y. Zong, *Multi-cue fusion for emotion recognition in the wild*, Neurocomputing, vol. 309, Elsevier B.V., p. 27-35, 2018.

[30] L. Chao, J. Tao, M. Yang, Y. Li and Z. Wen, *Multi-scale temporal modeling for dimensional emotion recognition in video*, AVEC 2014 - Proceedings of the 4th International Workshop on Audio/Visual Emotion Challenge, Workshop of MM 2014, Association for Computing Machinery, Inc, p. 11-18, 2014.

[31] C. Shan, S. Gong and P. W. McOwan, *Facial expression recognition based on Local Binary Patterns: A comprehensive study*, Image and Vision Computing, Elsevier Ltd, p. 803-816, 2009.

[32] T. Kanade, J. Cohn and Y. Tian, *Comprehensive database for facial expression analysis*, in: IEEE International Conference on Automatic Face & Gesture Recognition (FG), 2000.

[33] S. E. Kahou, X. Bouthillier, P. Lamblin, C. Gulcehre, V. Michalski, K. Konda, S. Jean et. al, *EmoNets: Multimodal deep learning approaches for emotion recognition in video*, arXiv 1503.01800, 2015.

[34] S. Tivatansakul, M. Ohkura, S. Puangpontip and T. Achalakul, *Emotional healthcare system: Emotion detection by facial expressions using Japanese database*, 2014 6th Computer Science and Electronic Engineering Conference, CEEC 2014 - Conference Proceedings, 2014.

[35] W. Zhang, D. Zhao, X. Chen and Y. Zhang, *Deep learning based emotion recognition from Chinese speech*, Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), Springer Verlag, 2016.