Video-based Emotion Detection Analyzing Facial Expressions and Contactless Vital Signs for Psychosomatic Monitoring

Hayette Hadjar[a], Binh Vu[a], Dennis Maier[a], Gwendolyn Mayer[b], Paul Mc Kevitt[c], and Matthias Hemmje[d]

a. *University of Hagen, Faculty of Mathematics and Computer Science, Hagen, Germany.*
b. *Heidelberg University, Department of Internal Medicine II, General Internal Medicine and Psychosomatics, Heidelberg, Germany.*
c. *Ulster University, Derry/Londonderry, Northern Ireland.*
d. *Research Institute for Telecommunication and Cooperation, Dortmund, Germany.*

Abstract

Recently, automatic emotion recognition research is gaining attention in computer science. This paper outlines research on Affective Computing in the context of the Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) project for remote home healthcare applications. This includes: (1) developing the system architecture for monitoring emotions and vital signs using a simple camera, and (2) recognition and visualization of emotions and of physiological signals to synthesize patients' psychosomatic data for health-care providers. In this exemplar prototypical implementation we employ Convolutional Neural Networks (CNNs) and remote PhotoPlethysmoGraphy (rPPG) methods for recognition of psychosomatic states during patient monitoring.

Keywords

Continuous Emotion Recognition, Facial Expressions, Remote Photoplethysmography (rPPG), Contactless Physiological Signals, Affective Computing

1. Introduction and Motivation

Psychosomatic medicine is becoming a key medical specialty that targets a deeper understanding of the physical, emotional and social causes for a disease [1]. Relevance of the corresponding treatment of patients has been significantly magnified by the coronavirus pandemic [2] and it has boosted the practice of telemedicine [3]. It appears tele-home health care settings give a rich landscape for research and application of affective computing to improve the quality of patient care [4]. This work has been developed in the context of the Sensor Enabled Affective Computing for Enhancing Medical Care (SenseCare) project [5]. SenseCare is a research and innovation project which was initially funded by the European Union [6] [7]. SenseCare focuses on developing a number of input interfaces for specific sensory devices, e.g., cameras and wearable sensors from the Internet of Things. Several analysis methods for emotion recognition within the web platform of SenseCare have already been developed. Three analysis methods are currently available on the SenseCare KM-EP (Knowledge Management Ecosystem Platform): (1) analysis based on Support Vector Machines according (see Healy et al. [8]), (2) Artificial Neural Networks (see Maier [9]), and (3) Convolutional Neural Networks with TensorFlow (see Hadjar et al. [10]). In the world of Big Data, data visualization tools and technologies are essential to analyze large sets of information and to make data-driven decisions. The facial appearance of patients may indeed give diagnostic clues to maladies, severity of diseases, and their vital parameters [11]. When it comes to very large and complex

CERC 2021: Collaborative European Research Conference, September 09–10, 2021, Cork, Ireland
hayette.hadjar@fernuni-hagen.de (H. Hadjar); binh.vu@fernuni-hagen.de (B. Vu); dennis.maier@fernuni-hagen.de (D. Maier) ; gwendolyn.mayer@med.uni-heidelberg.de (G. Mayer) ; p.mckevitt@ulster.ac.uk (P. McKevitt) ; mhemmje@ftk.de (M. Hemmje) ; http://www.paulmckevitt.com/ (P. McKevitt);
0000-0001-9540-6473 (H. Hadjar); 0000-0001-9715-1590 (P. McKevitt);

datasets, the functionality of the data visualization library at play is also important for effective insights. In this preliminary study, we are primarily interested in the following two goals:

- Visualization and perception of all collected audio, video, and emotion feature annotations or graphical representation of emotions over time, in order to make optimal healthcare decisions.
- Extraction of physiological signals (pulse rate, respiratory rate) from the camera, detection of potential changes of various emotions, and characterization of these changes.

The remainder of this paper is organized as follows. Section 2 discusses the state of the art of video-based facial expression emotion recognition. Section 3 details the conceptual design of an architecture for recognition and visualization of facial expressions and contactless vital signs. Section 4 describes the implementation of the proposed system, and finally we conclude and discuss future work in Section 5.

2. State of the Art and Related Work

Machine learning algorithms are applied to a wide variety of areas, such as spam filtering, speech recognition, face recognition, document classification and the processing of natural language. Classification is one of the most common areas of machine learning application. Recently, Video Facial Emotion Recognition research has received focus of attention in the computer vision community. In the task of emotion recognition, varied types of input data are employed such as: facial expressions, speech, physiological parameters, and body gestures. There are many approaches for detection of emotion from facial expressions, e.g., the work of Michel Healy [8], where an emotion detection system is described based on a video feed in real-time, and employs a machine learning Support Vector Machine (SVM) to provide quick and reliable classification. Features employed in [8] are 68-point facial landmarks. The application has been trained to detect 6 different emotions by monitoring changes in facial expressions. The work of Dennis Maier [9] currently uses neural networks via TensorFlow to train image features and then achieves classification through fully connected neural layers. The advantage of image features over facial landmarks is the larger information space, where the spatial formation of landmarks gives a viable method for analyzing facial expressions. However, this is also accompanied by a higher computing power requirement. The structure provides for an outsourced classification service that runs on a server with a GPU. Images of faces are brought to the service in real time, which can perform a classification within a few milliseconds. In future, this approach will be extended to include text and audio features and conversation context to boost accuracy. Another approach uses Convolutional Neural Networks with TensorFlow [12]. An example for using TensorFlow.js with the SenseCare KM-EP is discussed in [8], which deploys a web browser and Node Server [13].

Furthermore, Speech Emotion Recognition is an important field employed by numerous multimodal sentiment analysis systems, e.g., robotics, security, automated identification, and language translation [14]. The MENHIR project [15] aims to investigate conversational technologies to promote mental health and assist people with mental ill health in managing their conditions. In the context of MENHIR, a new emotion-recognition system based on speech is being developed [16]. Physiological signs are detected such as facial electromyography, facial color patterns, blood volume pulse, and galvanic skin response. The research project "EEG-based Emotion Recognition" [17] leads in recognizing emotion from brain signals measured with the Bra Inquiry EEG PET device. Furthermore, the work in [18] implements a wearable sensing system for effective recognition of user emotional states, by employing heart rate, body temperature, and galvanic skin response sensors. The proposed system achieves up to 97% recognition accuracy when it adopts the k-nearest neighbor (KNN) classifier. For contactless vital signs, several methods are available for camera-based Heart Rate (HR), Heart Rate Variability (HRV), and Respiration Rate (RR) detection such as Remote PhotoPlethysmography (rPPG) where [19] describes a popular technique of non-contact measure HR from facial videos, and [20] describes how RR works. Recently, studies on deep learning based rPPG methods have been introduced such as [21] and [22]. Eulerian Video Magnification [23] applies spatial decomposition to the input standard video sequences, where amplification reveals hidden information in resulting signals. Other research employs Convolutional Neural Networks for remote pulse rate measurement and mapping from facial video, such as the DeepPhys Convolutional Neural

Network [24] and 3D Convolutional Neural Network [25]. Furthermore, several approaches exist for body gesture recognition, e.g., Noroozi et al. [26]. This survey classifies body language into 6 basic emotions: fear, anger, sadness, surprise, happiness, and disgust. Another study [27] introduces deep learning models for multivariate time series, from vehicle control to gesture recognition and generation.

3. Conceptual Architecture

The processing of large data streams from audio-video recordings (real-time and offline) to data visualization follows the architecture of the Affective Computing Strata (AC-Strata) model [28] where the S-Strata relates to an individual mode of affective monitoring. The S-Strata Information model indicates that under any AC analytical method, one, or any number of specific sensors of embedded devices, may participate. The S-Strata model describes the time aspects of the sensors that are integrated in a specific manner and also stipulates contextual and manual representations dealing with both internal and external features. Accordingly, the conceptual architecture of our SenseCare affective computing platform based on audio- and video-based emotion recognition within the SenseCare KM-EP is detailed in Figure 1.

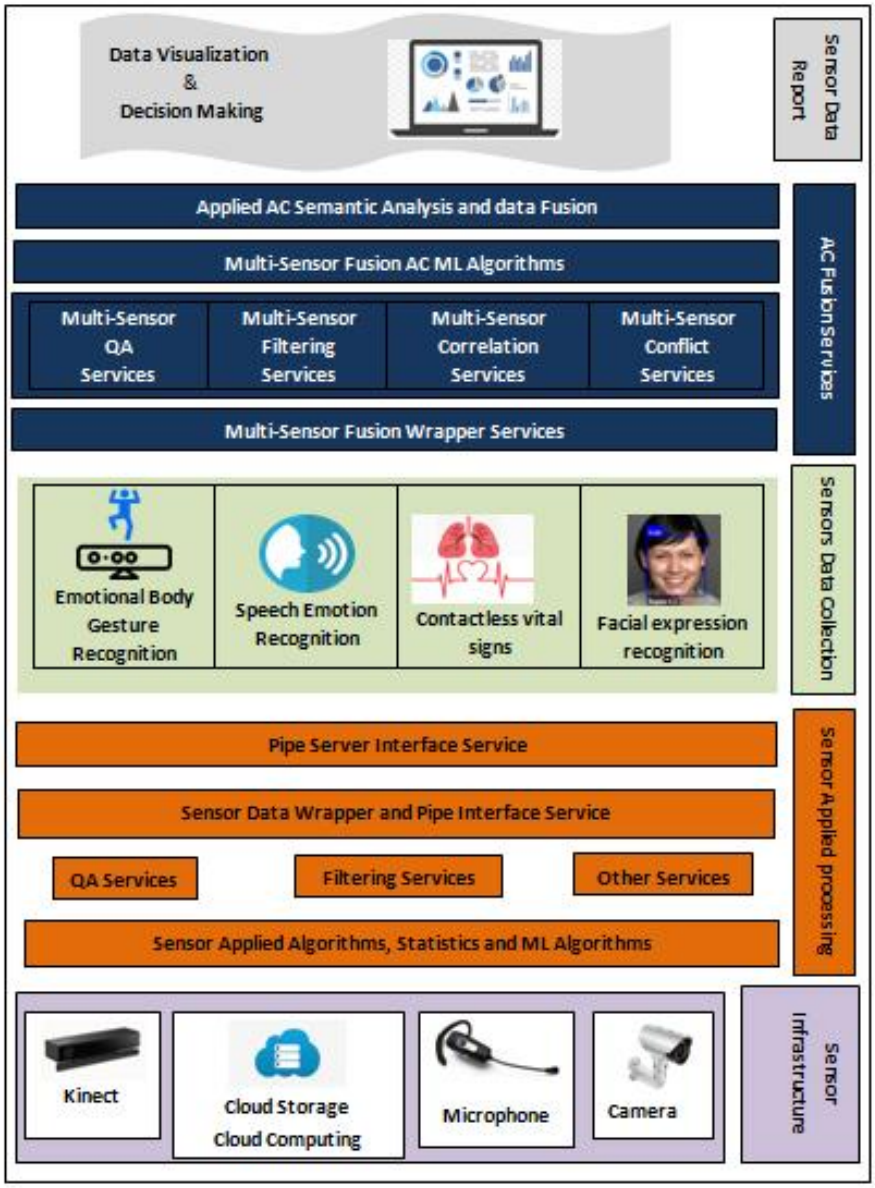

Figure 1: Conceptual Architecture of the SenseCare Affective Computing Platform

The conceptual architecture design is structured as follows. First of all, in the sensor & infrastructure area, sensors used are: (1) Camera (for Facial Expression Recognition, and Contactless

Vital Signs [HR, RR, HRV]), (2) Microphone (for Speech-based Emotion Recognition), and (3) Kinect (for Body Gesture emotion recognition). Furthermore, cloud storage & cloud computing is employed here as an infrastructure for data processing and storage. Secondly, processing applied to sensor data includes facial expression-based emotion recognition employing analysis methods discussed above, i.e., Support Vector Machines (Healy et al. [8]), Artificial Neural Networks (Maier [9]), and Convolutional Neural Networks with TensorFlow (Hadjar et al. [10]), deployed in a web browser and Node Server [13]. Finally, sensor data feature collection includes the following 4 types of Affective Computing data features: (1) Facial expression-based emotion recognition, (2) Speech-based emotion recognition, (3) Contactless Vital Signs (HR, RR, HRV), and (4) Body-Gesture based emotion recognition. On the higher level of the architecture, the AC Fusion services are responsible for the fusion of the collected AC data features. On top of this data fusion layer the architecture user interface layer generates advanced services for sensor data reporting including data visualization and decision making support available to end users.

4. Implementation

The prototypical implementation of this research concentrates on extracting emotion-related features from images of the human face, ideally in real-time, from a single sensor which is the camera. Hence, the current state of implementation of AC sensor data collection is the facial-expression based emotion recognition and the vital signs recognition. An implementation of a prototypical module that collects patients' facial expression and corresponding emotion data in real-time was already described in detail in our previous work [10]. The API allowed monitoring patients during treatment sessions, or at home for cases of patients with or at risk for a mental disorder. The software classifies emotions into 7 categories: happy, sad, angry, disgusted, fearful, neutral, and surprised, as determined by Paul Ekman [29]. The prototype employs deep learning in browsers using JavaScript. In the case of real-time video-based emotion analysis, the SenseCare KM-EP's Emotion Detection API stores the most significant emotion detected from the video every 500 milliseconds into a database. The face expression recognition model employs depthwise separable convolutions and densely connected blocks. The Perceptron model is a linear transformation, where the activation function makes it possible for the model to approximate non-linear functions (e.g. Softmax, Sigmoid). We have chosen line chart graphs based on Chart.js [30] to visualize stored emotions in the database during time periods of, e.g., the last 24h, the last 3 days, the last week, and the last month. Graphs showing the score distributions of emotions over time are given in Figures 2-5 below.

Figure 2: Emotions detected (last 24 hours)

Figure 3: Emotions detected (last 3 days)

Figure 4: Emotions detected (last week)

Figure 5: Emotions detected (last month)

Data acquired over several days (last 30 days) makes it possible to build a comprehensive picture compared to that acquired over a single day (last 24h). This data is potentially useful for analyzing psychosomatic states.

The development of video-based contactless sensing of vital signs yields opportunities for scalable physiological monitoring. For the measurement of stress, Heart Rate Variability is the focus of much work. Experimentation with a video-based pulse rate monitoring implementation in browser-based rPPG employing the Viola-Jones algorithm [31] for defining the Region of Interest and Haar-like features [32] for face detection was tested as shown at the following URL: https://studev4.fernuni-hagen.de:20286/

In respect of Data Fusion, visual facial expression analysis may in many cases be insufficient for emotion recognition [33]. Hence, we suggest in this approach the fusion of vital factors with the analysis of facial expressions to give greater precision of predicted results. The multimodal approach has been proposed for enabling more reliable recognition with minimal ambiguities that can arise

whilst using a single data channel. Bimodal emotion analysis of camera-based vital sign and facial expression recognition is shown in Figure 6.

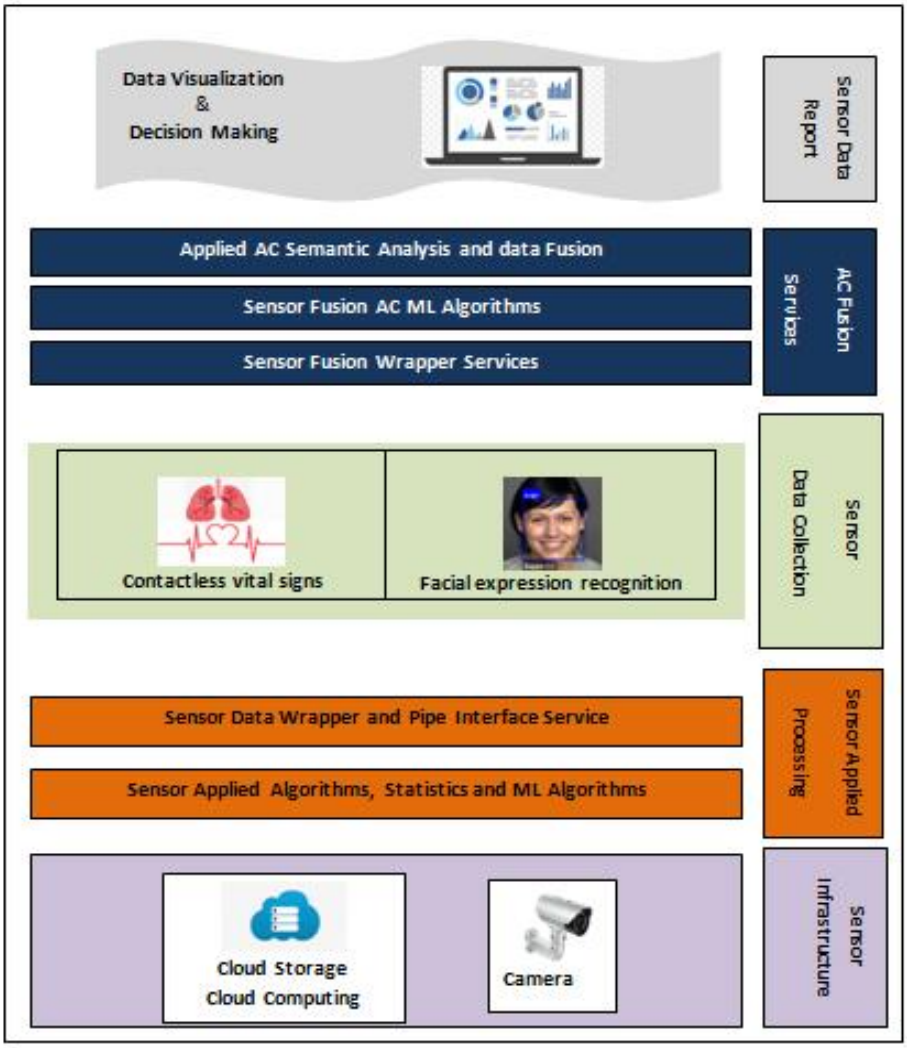

Figure 6: Multimodal Emotion Recognition based on Video

In this context, we must also handle the real-time tasks of acquiring and merging information and system performance.

5. Conclusion and Future Work

In this study, an approach of detection and recognition of emotions from the camera is proposed; bimodal systems will analyze facial emotions using deep learning and other algorithms by merging the two collected emotions. The current system recognizes emotions from facial expressions using Convolutional Neural Networks with Tensorflow.js, and heart pulse using the rPPG technique with the Viola-Jones algorithm, and finally displays results in the web browser. The advantages of the proposed system are:

- All the processing of the AI facial emotion analysis is done on the client machine, and no data such as real-time video is sent to the server.
- It is low cost with more comfort for the patient (no contact with skin).
- The system is accessible from any device with a camera.
- The system enables displaying a graphical result in a variety of devices, such as a Smartphone.
- The Node development environment supports a variety of popular libraries such as D3.js and Ggplot2, which can be included later, and enables collaborative visualization.

In this paper, we introduced a conceptual architecture of an affective computing platform supporting audio- and video-based emotion recognition within the SenseCare KM-EP. A new approach is proposed which includes facial expression recognition and contactless vital signs to provide an initial psychosomatic monitoring for the SenseCare KM-EP. The proposed system offers computer diagnostics and evaluation of emotions supporting diagnosis and treatment of psychosomatic illnesses. Our next steps are: (1) fusion of AC data and global visualization of emotions collected based on video and audio, (2) investigate the treatment of anxiety using eye-tracking and heart rate variability, and test and build different datasets and models, (3) develop a mobile application to offer a variety of choices for our users and improve the accessibility of our system, (4) investigate user data privacy, and (5) develop other systems with ability to detect different types of emotions and potential diseases. In future work, it will be important to evaluate the bimodal system with real patients to improve their performance.

6. References

[1] K. Fritzsche, S. H. McDaniel, and M. Wirsching, Eds., Psychosomatic Medicine. 2020.

[2] Yu-Tao Xiang, "Timely mental health care for the 2019 novel coronavirus outbreak is urgently needed, Lancet Psychiatry, vol. 7, no. 3, pp. 228–229, 2020.

[3] J. Torous, K. Jän Myrick, N. Rauseo-Ricupero, and J. Firth, Digital Mental Health and COVID-19: Using Technology Today to Accelerate the Curve on Access and Quality Tomorrow,JMIR Ment. Heal., vol. 7, no. 3, p. e18848, Mar. 2020, doi: 10.2196/18848.

[4] C. L. Lisetti and C. LeRouge, Affective computing in tele-home health: design science possibilities in recognition of adoption and diffusion issues, in HICSS 2004, 37,th IEEE Hawaii International Conference on System Sciences, January 5-8, 2004, Hawaii, USA, 2004.

[5] F. Engel et al., SenseCare: Towards an Experimental Platform for Home-Based, Visualisation of Emotional States of People with Dementia, in Advanced Visual Interfaces. Supporting Big Data Applications, 2016, pp. 63–74.

[6] Sensor Enabled Affective Computing for Enhancing Medical Care | SenseCare Project | H2020 | CORDIS | European Commission.URL: https://cordis.europa.eu/project/id/690862/fr.

[7] SenseCare: Sensor Enabled Affective Computing for Enhancing Medical Care | FTK – Research Institute for Telecommunication and Cooperation.URL: https://www.ftk.de/en/projects/senscare.

[8] M. Healy, R. Donovan, P. Walsh, and H. Zheng, A Machine Learning Emotion Detection Platform to Support Affective Well Being, in 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2018, pp. 2694–2700, doi: 10.1109/BIBM.2018.8621562.

[9] D. Maier, Analysis of technical drawings by using deep learning, Master's thesis, Department of Computer Science, Hochschule Mannheim University, Germany, 2019.

[10] H. Hadjar, Video-based automated emotional monitoring in mental health care supported by a generic patient data management system, 2020.

[11] M. Leo and G. M. Farinella, Preface in Computer Vision for Assistive Healthcare, M. Leo and G. M. Farinella, Eds. Academic Press, 2018, pp. xxi–xxiii.

[12] M. N. A. Wisal Hashim Abdulsalam, Rafah Shihab Alhamdani, Facial Emotion Recognition from Videos Using Deep Convolutional Neural Networks, Int. J. Mach. Learn. Comput., vol. 9, no. 1, pp. 14–19, 2019.

[13] Node.js, URL: https://nodejs.org/en/

[14] M. El Ayadi, M. S. Kamel, and F. Karray, Survey on speech emotion recognition: Features, classification schemes, and databases, Pattern Recognit., vol. 44, no. 3, pp. 572–587, 2011, doi: https://doi.org/10.1016/j.patcog.2010.09.020.

[15] Mental health monitoring through interactive conversations | MENHIR Project | H2020 | CORDIS | European Commission. URL: https://cordis.europa.eu/project/id/823907.

[16] B. Vu., A Content and Knowledge Management System Supporting Emotion Detection from Speech BT - Conversational Dialogue Systems for the Next Decade, L. F. D'Haro, Z. Callejas, and S. Nakamura, Eds. Singapore: Springer Singapore, 2021, pp. 369–378.

[17] D. Bos, EEG-based Emotion Recognition The Influence of Visual and Auditory Stimuli, 2007.
[18] B. Myroniv, C.-W. Wu, Y. Ren, and Y.-C. Tseng, Analysis of Users' Emotions Through Physiology, in Genetic and Evolutionary Computing, 2018, pp. 136–143.
[19] X. Ma, D. P. Tobón, and A. El Saddik, Remote Photoplethysmography (rPPG) for Contactless Heart Rate Monitoring Using a Single Monochrome and Color Camera, in Smart Multimedia, 2020, pp. 248–262.
[20] M. Chen, Q. Zhu, H. Zhang, M. Wu, and Q. Wang, Respiratory Rate Estimation from Face Videos, in 2019 IEEE EMBS International Conference on Biomedical Health Informatics (BHI), 2019, pp. 1–4, doi: 10.1109/BHI.2019.8834499.
[21] R. Song, S. Zhang, C. Li, Y. Zhang, J. Cheng, and X. Chen, Heart Rate Estimation From Facial Videos Using a Spatiotemporal Representation With Convolutional Neural Networks, IEEE Trans. Instrum. Meas., vol. 69, no. 10, pp. 7411–7421, 2020, doi: 10.1109/TIM.2020.2984168.
[22] Q. Zhan, W. Wang, and G. de Haan, Analysis of CNN-based remote-PPG to understand limitations and sensitivities, Biomed. Opt. Express, vol. 11, no. 3, pp. 1268–1283, Mar. 2020, doi: 10.1364/BOE.382637.
[23] H.-Y. Wu, M. Rubinstein, E. Shih, J. Guttag, F. Durand, and W. Freeman, Eulerian Video Magnification for Revealing Subtle Changes in the World, ACM Trans. Graph. - TOG, vol. 31, 2012, doi: 10.1145/2185520.2185561.
[24] W. Chen and D. McDuff, DeepPhys: Video-Based Physiological Measurement Using Convolutional Attention Networks, in Computer Vision -- ECCV 2018, 2018, pp. 356–373.
[25] F. Bousefsaf, A. Pruski, and C. Maaoui, 3D Convolutional Neural Networks for Remote Pulse Rate Measurement and Mapping from Facial Video, Appl. Sci., vol. 9, p. 4364, 2019, doi: 10.3390/app9204364.
[26] F. Noroozi, C. Corneanu, D. Kaminska, T. Sapinski, S. Escalera, and G. Anbarjafari, Survey on Emotional Body Gesture Recognition, IEEE Trans. Affect. Comput., vol. 12, no. 02, pp. 505–523, 2021, doi: 10.1109/TAFFC.2018.2874986.
[27] G. Devineau, Deep learning for multivariate time series : from vehicle control to gesture recognition and generation, Université Paris sciences et lettres, 2020.
[28] A. Keary, Affective Computing for Emotion Detection using Vision and Wearable Sensors, 2018.
[29] P. Ekman and G. Yamey, Emotions revealed: recognising facial expressions: in the first of two articles on how recognising faces and feelings can help you communicate, Paul Ekman discusses how recognising emotions can benefit you in your professional life, Student BMJ, vol. 12, pp. 140–142, 2004.
[30] Chart.js | Open source HTML5 Charts for your website.URL: https://www.chartjs.org/
[31] P. Viola and M. Jones, Rapid object detection using a boosted cascade of simple features, in Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001, 2001, vol. 1, pp. I–I, doi: 10.1109/CVPR.2001.990517.
[32] T. Mita, T. Kaneko, and O. Hori, Joint Haar-like features for face detection, in Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, 2005, vol. 2, pp. 1619-1626 Vol. 2, doi: 10.1109/ICCV.2005.129.
[33] A. Kwasniewska, J. Ruminski, M. Szankin, and K. Czuszynski, Remote Estimation of Video-Based Vital Signs in Emotion Invocation Studies, in Conference proceedings: Annual International Conference of the IEEE Engineering in Medicine and Biology Society. IEEE Engineering in Medicine and Biology Society. Conference, 2018, vol. 2018, pp. 4872–4876, doi: 10.1109/EMBC.2018.8513423.