Ana I. Pereira · Andrej Košir ·
Florbela P. Fernandes · Maria F. Pacheco ·
João P. Teixeira · Rui P. Lopes (Eds.)

Communications in Computer and Information Science 1754

Optimization, Learning Algorithms and Applications

Second International Conference, OL2A 2022
Póvoa de Varzim, Portugal, October 24–25, 2022
Proceedings

Springer

OL2A
INTERNATIONAL
CONFERENCE

Communications in Computer and Information Science 1754

Ana I. Pereira · Andrej Košir ·
Florbela P. Fernandes · Maria F. Pacheco ·
João P. Teixeira · Rui P. Lopes (Eds.)

Optimization, Learning Algorithms and Applications

Second International Conference, OL2A 2022
Póvoa de Varzim, Portugal, October 24–25, 2022
Proceedings

Springer

Editors
Ana I. Pereira
Instituto Politécnico de Bragança
Bragança, Portugal

Andrej Košir
University of Ljubljana
Ljubljana, Slovenia

Florbela P. Fernandes
Instituto Politécnico de Bragança
Bragança, Portugal

Maria F. Pacheco
Instituto Politécnico de Bragança
Bragança, Portugal

João P. Teixeira
Instituto Politécnico de Bragança
Bragança, Portugal

Rui P. Lopes
Instituto Politécnico de Bragança
Bragança, Portugal

ISSN 1865-0929 ISSN 1865-0937 (electronic)
Communications in Computer and Information Science
ISBN 978-3-031-23235-0 ISBN 978-3-031-23236-7 (eBook)
https://doi.org/10.1007/978-3-031-23236-7

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Preface

This CCIS volume 1754 contains the refereed proceedings of the Second International Conference on Optimization, Learning Algorithms and Applications (OL2A 2022), a hybrid event held during October 24–25, 2022.

OL2A 2022 provided a space for the research community on optimization and learning to get together and share the latest developments, trends, and techniques, as well as to develop new paths and collaborations. The conference had more than three hundred participants in an online and face-to-face environment throughout two days, discussing topics associated with optimization and learning, such as state-of-the-art applications related to multi-objective optimization, optimization for machine learning, robotics, health informatics, data analysis, optimization and learning under uncertainty, and Industry 4.0.

Five special sessions were organized under the following topics: Trends in Engineering Education, Optimization in Control Systems Design, Measurements with the Internet of Things, Advances and Optimization in Cyber-Physical Systems, and Computer Vision Based on Learning Algorithms. The OL2A 2022 program included presentations of 56 accepted papers. All papers were carefully reviewed and selected from 145 submissions in an single-blind process. All the reviews were carefully carried out by a scientific committee of 102 qualified researchers from 21 countries, with each submission receiving at least 3 reviews.

We would like to thank everyone who helped to make OL2A 2022 a success and hope that you enjoy reading this volume.

October 2022

Ana I. Pereira
Andrej Košir
Florbela P. Fernandes
Maria F. Pacheco
João P. Teixeira
Rui P. Lopes

Organization

General Chairs

Ana I. Pereira	Polytechnic Institute of Bragança, Portugal
Andrej Košir	University of Ljubljana, Slovenia

Program Committee Chairs

Florbela P. Fernandes	Polytechnic Institute of Bragança, Portugal
Maria F. Pacheco	Polytechnic Institute of Bragança, Portugal
João P. Teixeira	Polytechnic Institute of Bragança, Portugal
Rui P. Lopes	Polytechnic Institute of Bragança, Portugal

Special Session Chairs

João P. Coelho	Polytechnic Institute of Bragança, Portugal
Luca Oneto	University of Genoa, Italy

Technology Chairs

Alexandre Douplik	Ryerson University, Canada
Paulo Alves	Polytechnic Institute of Bragança, Portugal

Local Organizing Chairs

José Lima	Polytechnic Institute of Bragança, Portugal

Program Committee

Ana I. Pereira	Polytechnic Institute of Bragança, Portugal
Abeer Alsadoon	Charles Sturt University, Australia
Ala' Khalifeh	German Jordanian University, Jordan
Alexandre Douplik	Ryerson University, Canada
Ana Maria A. C. Rocha	University of Minho, Portugal
Ana Paula Teixeira	University of Trás-os-Montes and Alto Douro, Portugal
André Pinz Borges	Federal University of Technology – Paraná, Brazil
André R. da Cruz	Federal Center for Technological Education of Minas Gerais, Brazil

Andrej Košir	University of Ljubljana, Slovenia
Arnaldo Cândido Júnior	Federal University of Technology – Paraná, Brazil
António J. Sánchez-Salmerón	Universitat Politécnica de Valencia, Spain
António Valente	Universidade de Trás-Os-Montes e Alto Douro, Portugal
Bilal Ahmad	University of Warwick, UK
Bruno Bispo	Federal University of Santa Catarina, Brazil
Carlos Henrique Alves	CEFET/RJ, Brazil
Carmen Galé	University of Zaragoza, Spain
B. Rajesh Kanna	Vellore Institute of Technology, India
Carolina Gil Marcelino	Federal University of Rio de Janeiro, Brazil
Christopher E. Izquierdo	University of Laguna, Spain
C. Sweetlin Hemalatha	Vellore Institute of Technology, India
Damir Vrančić	Jozef Stefan Institute, Slovenia
Daiva Petkeviciute	Kaunas University of Technology, Lithuania
Dhiah Abou-Tair	German Jordanian University, Jordan
Diego Brandão	CEFET/RJ, Brazil
Dimitris Glotsos	University of West Attica, Greece
Diamantino Silva Freitas	University of Porto, Portugal
Eduardo Vinicius Kuhn	Federal University of Technology – Paraná, Brazil
Elizabeth Fialho Wanner	Federal Center for Technological Education of Minas Gerais, Brazil
Elaine Mosconi	Université de Sherbrooke, Canada
Esteban Clua	Federal Fluminense University, Brazil
Eric Rogers	University of Southampton, UK
Felipe N. Martins	Hanze University of Applied Sciences, Netherlands
Florbela P. Fernandes	Polytechnic Institute of Bragança, Portugal
Florentino F. Riverola	University of Vigo, Spain
Gaukhar Muratova	Dulaty University, Kazakhstan
Gediminas Daukšys	Kauno Technikos Kolegija, Lithuania
Glaucia Maria Bressan	Federal University of Technology – Paraná, Brazil
Glotsos Dimitris	University of West Attica, Greece
Humberto Rocha	University of Coimbra, Portugal
J. Marcos Moreno Veja	University of Laguna, Spain
João Paulo Carmo	University of São Paulo, Brazil
João P. Teixeira	Polytechnic Institute of Bragança, Portugal
Jorge Igual	Universidad Politécnica de Valencia, Spain
José Boaventura-Cunha	University of Trás-os-Montes and Alto Douro, Portugal
José Lima	Polytechnic Institute of Bragança, Portugal

Joseane Pontes	Federal University of Technology – Ponta Grossa, Brazil
Juan Alberto G. Esteban	Universidad de Salamanca, Spain
Juan A. Méndez Pérez	University of Laguna, Spain
Juani Lopéz Redondo	University of Almeria, Spain
Julio Cesar Nievola	Pontifícia Universidade Católica do Paraná, Brazil
João Paulo Coelho	Polytechnic Institute of Bragança, Portugal
Jorge Ribeiro	Polytechnic Institute of Viana do Castelo, Portugal
José Ramos	NOVA University Lisbon, Portugal
Kristina Sutiene	Kaunas University of Technology, Lithuania
Lidia Sánchez	University of León, Spain
Lino Costa	University of Minho, Portugal
Luis A. De Santa-Eulalia	Université de Sherbrooke, Canada
Luís Coelho	Polytecnhic Institute of Porto, Portugal
Luca Oneto	University of Genoa, Italy
Luca Spalazzi	Marche Polytechnical University, Italy
Maria F. Pacheco	Polytechnic Institute of Bragança, Portugal
Manuel Castejón Limas	University of León, Spain
Marc Jungers	Université de Lorraine, France
Marco A. S. Teixeira	Universidade Tecnológica Federal do Paraná, Brazil
Maria do R. de Pinho	University of Porto, Portugal
Marco A. Wehrmeister	Federal University of Technology – Paraná, Brazil
Markus Vincze	TU Wien, Austria
Martin Hering-Bertram	Hochschule Bremen, Germany
Mikulas Huba	Slovak University of Technology in Bratislava, Slovakia
Michał Podpora	Opole University of Technology, Poland
Miguel Ángel Prada	University of León, Spain
Nicolae Cleju	Technical University of Iasi, Romania
Paulo Lopes dos Santos	University of Porto, Portugal
Paulo Alves	Polytechnic Institute of Bragança, Portugal
Paulo Leitão	Polytechnic Institute of Bragança, Portugal
Paulo Moura Oliveira	University of Trás-os-Montes and Alto Douro, Portugal
Pavel Pakshin	Nizhny Novgorod State Technical University, Russia
Pedro Luiz de P. Filho	Federal University – Paraná, Brazil
Pedro Miguel Rodrigues	Catholic University of Portugal, Portugal
Pedro Morais	Polytechnic Institute of Cávado e Ave, Portugal
Pedro Pinto	Polytechnic Institute of Viana do Castelo, Portugal

Roberto M. de Souza	Federal University of Technology – Paraná, Brazil
Rui P. Lopes	Polytechnic Institute of Bragança, Portugal
Sabrina Šuman	Polytechnic of Rijeka, Croatia
Sani Rutz da Silva	Federal University of Technology – Paraná, Brazil
Santiago Torres Álvarez	University of Laguna, Spain
Sara Paiva	Polytechnic Institute of Viana do Castelo, Portugal
Shridhar Devamane	Global Academy of Technology, India
Sofia Rodrigues	Polytechnic Institute of Viana do Castelo, Portugal
Sławomir Stępień	Poznan University of Technology, Poland
Teresa P. Perdicoulis	University of Trás-os-Montes and Alto Douro, Portugal
Toma Roncevic	University of Split, Croatia
Uta Bohnebeck	Hochschule Bremen, Germany
Valeriana Naranjo-Ornedo	Universidad Politécnica de Valencia, Spain
Vivian Cremer Kalempa	Universidade Estadual de Santa Catarina, Brazil
Vitor Duarte dos Santos	NOVA University Lisbon, Portugal
Wynand Alkema	Hanze University of Applied Sciences, Netherlands
Wojciech Paszke	University of Zielona Gora, Poland
Wojciech Giernacki	Poznan University of Technology, Poland
Wolfgang Kastner	TU Wien, Austria

Classification of Facial Expressions Under Partial Occlusion for VR Games

Ana Sofia Figueiredo Rodrigues[1], Júlio Castro Lopes[1], Rui Pedro Lopes[1,2(✉)], and Luís F. Teixeira[3]

[1] Research Center in Digitalization and Intelligent Robotics (CeDRI), Instituto Politécnico de Bragança, Bragança, Portugal
a41737@alunos.ipb.pt, {juliolopes,rlopes}@ipb.pt
[2] Institute of Electronics and Informatics Engineering of Aveiro (IEETA), University of Aveiro, Aveiro, Portugal
[3] INESC TEC, Faculdade de Engenharia da Universidade do Porto, Porto, Portugal
luisft@fe.up.pt

Abstract. Facial expressions are one of the most common way to externalize our emotions. However, the same emotion can have different effects on the same person and has different effects on different people. Based on this, we developed a system capable of detecting the facial expressions of a person in real-time, occluding the eyes (simulating the use of virtual reality glasses). To estimate the position of the eyes, in order to occlude them, Multi-task Cascade Convolutional Neural Networks (MTCNN) were used. A residual network, a VGG, and the combination of both models, were used to perform the classification of 7 different types of facial expressions (Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral), classifying the occluded and non-occluded dataset. The combination of both models, achieved an accuracy of 64.9% for the occlusion dataset and 62.8% for no occlusion, using the FER-2013 dataset. The primary goal of this work was to evaluate the influence of occlusion, and the results show that the majority of the classification is done with the mouth and chin. Nevertheless, the results were far from the state-of-the-art, which is expect to be improved, mainly by adjusting the MTCNN.

Keywords: Facial expression recognition · Emotions · ResNet · VGG · MTCNN · Virtual reality

1 Introduction

Facial expression recognition is playing an increasingly critical role in several areas and applications. Beyond verbalization, non-verbal cues, such as facial emotions, play a vital role, providing clues in understanding and inferring the hidden emotional state of individuals. In fact, facial expressions are the most common way to externalize emotions and one of the most important means of communication in interpersonal relationships.

Humans have evolved to easily read the other people's facial expressions. Often these are expressed automatically without the transmitter even realizing

A. I. Pereira et al. (Eds.): OL2A 2022, CCIS 1754, pp. 804–819, 2022.
https://doi.org/10.1007/978-3-031-23236-7_55

that he is executing them [29]. Based on these signals, we adjust our behaviour and make choices according to our perception of the emotional state of the other person. In other words, the interaction may be conditioned by the mutual interpretation of the emotional state, derived from reading the other's facial expression (and complemented with body language, voice tone, and others).

Extending this possibility to computers, the automatic identification of human facial expressions introduces several benefits. It has the potential to develop better and more useful human-computer interaction, provide visually impaired with haptic clues regarding the expression of others [24], monitor the motivation of students in the classroom [25], among many other applications. Although difficult, it is constantly evolving with several proposed solutions by the scientific community [35].

The work described in this paper aims to investigate and evaluate Machine Learning (ML) algorithms to infer human emotional state through the classification of facial expressions. Specifically, two tasks are explored: face detection and expression recognition. The main objective of this work is to develop a system capable of evaluating humans facial expression in real-time and, eventually, use this information to assess the emotional state while playing a virtual reality game.

This work is developed within the GreenHealth project [17]. This project addresses the use of digital technologies, namely Virtual Reality (VR), Game-based Learning (GBL), Internet of Things (IoT), Artificial Intelligence (AI), and Advanced Data Analytics (ADA), in the development of innovative rehabilitation techniques, contributing to the promotion of well-being and health in a holistic perspective, focusing, mainly on mental health rehabilitation in people diagnosed on the schizophrenia spectrum. Together with patient's body posture analysis, physiological data (e.g. patient's heart rate), and performance in the game, facial expression recognition enables to automatically adapt the game to the individual characteristics of the player. Since the player is wearing VR goggles, the classification of expressions has to be performed with part of the face hidden, which introduces further challenges. The purpose of this paper is to study the impact of occlusion in the classification of facial expressions for the dynamic adaptation of the difficulty level in cognitive rehabiliation serious games [16].

The paper is organized as follows: Sect. 2 presents the related work, and Sect. 3 discusses the methodology behind this study. Section 4 describes the results, and Sect. 5 rounds up the paper with the conclusions, particularly analyzing the challenges associated to the facial expressions classification.

2 Related Work

Over the centuries there have been several attempts to organize emotions into several categories. Cicero [20], held that they should be divided into four categories: fear (*metus*), pain (*aegritudo*), lust (*libido*) and pleasure (*laetiti*). Darwin and Prodger [21], for example, have used twenty-two categories, including anxiety, joy, love, devotion, anger, helplessness, surprise, fear, among others.

More recently, Plutchik have introduced the wheel of emotions, with 8 primary dimensions of emotion, namely joy, trust, fear, surprise, sadness, disgust, anger, anticipation [8].

The sole analysis of facial expressions is not enough to assess a person's sentiment. However, 55% of information is communicated by facial expressions, 38% by other gestures and signals (such as voice and sound) and 7% by spoken language [19]. Based on this, facial expression recognition account as an important factor for human sentiment recognition. With constantly improving of the computing power, jointly with the development of big data processing technology and algorithms, the automatic classification of facial expressions have been growing in accuracy and interest.

Devries et al. [7], demonstrated that a system trained to reason about facial geometry while recognizing expressions outperforms one trained solely to recognize expressions. Since facial landmark prediction has many publicly implementations available, the authors decided to use Zhu and Ramanan's facial landmark detector, which uses coordinates for 68 reference points per face. All of these points are used to outline facial features like the mouth, nose, eyes, and eyebrows. The authors simplified the problem by focusing on the most expressive features of humans facial expression: the eyebrows and mouth (rough reference points). Each of the positions of the left brow, right brow, and mouth were represented by a binary mask image.

The authors have used a Convolutional Neural Network (CNN) that was based on the winning architecture of the 2013 International Conference on Machine Learning (ICML) Facial Expression Recognition Competition [28]. They opted for a CNN with 3 convolutional layers fully connected, each with a ReLU activation function and max pooling. The network's output consisted of three binary output maps (one for each of the reference locations), and it was thus in charge of modeling the location and shape of each of its features. Two datasets were used in their work: ICML Dataset [10] and Toronto Face Database (TFD) [26]. The ICML consists of 28709 48×48 training images, each with 7 labels and 7177 test images. Since the images were taken from a wild environment, the faces have different orientations and are not always facing forward. In this dataset, the authors evaluated 2 different techniques: CNN with 3 convolutional layers fully connected and a Multi-task CNN. The TFD is also made up of 48×48 images and 7 labels, however, all faces are looking directly at the camera. It contains 4178 images, 70% of the images were used for training, 10% for validation, and 20% for testing. The authors used the same evaluation techniques applied in ICML dataset.

Baltrusaitis et al. [3] use the OpenFace facial behavior analyses pipeline, which includes the following algorithms: facial landmark detection, head pose tracking, eye gaze and facial Action Units (AUs). For the facial landmark detection and tracking, the authors have used the Conditional Local Neural Fields (CLNF), with 2 main components: Point Distribution Model (PDM) and patch experts. The PDM was been trained in 2 datasets, the Labeled Face Parts in the Wild (LFPW) and Helen. The CLNF patch experts was been trained in 3 datasets, including the ones used in the PDM and also the Multi-Pie dataset [11].

In order to estimate the head pose, information about the position of the head (translation and orientation) was extracted, as well as the detection of facial landmarks. This information was obtained thanks to the CLNF, which internally uses a 3D facial landmark representation and projects it to the image using orthographic camera projection. To train their model the authors have used the Mpiigaze [34] dataset.

Regarding the eye gaze estimation, CLNF and PDM were also used, since they allow detecting reference points of the eye region, such as the eyelids, iris and pupil. The PDM was trained with the Syntheseyes [30] dataset. The information obtained by the CLNF was used to compute the eye gaze vector individually for each eye (lightning is fired from the camera's source through the center of the pupil in the image plan, and its intersection is computed to determine the pupil's location in 3D camera coordinates).

Finally Openface predicts AU presence using a linear Support Vector Machine (SVM) kernel and AU intensity using a linear Support Vector Regression (SVR) kernel.

Loizou [15] proposed and evaluated a system for analyzing automated speech signals and images for seven different human emotions: normal, happy, sad, dislike, fear, anger, and surprise. Voice and image recordings of more than 70000 people aged twenty to seventy-four years old were organized. The authors have used multi-classification models to select the features that identify the seven emotions through an SVM with a 10-fold validation, using a Gaussian Radial Basis Function (RBF) with $c = 1$ and $\gamma = 0.01$. Statistically, a Correct Classification (CC) score of 93% was obtained.

Mindlink-Eumpy [14], an open-source toolbox, was designed to detect emotions by integrating information from Electrocardiogram (EEG) and facial expressions. First, a set of tools was used to automatically collect physiological data, which was then used to analyze user facial expressions and EEG data. Regarding the analysis of user's facial expression, they used a multitask CNN pre-trained with the FER2013 dataset [10].

On the surface, the idea of using a pre-trained CNN, is to transfer the labeled data or knowledge from some domains (previously performed tasks - Source Tasks), to help the ML algorithm to perform better in the domain of interest. Regarding the EEG analysis, MindLink-Eumpy [14] uses two different algorithms: SVM and Long Short-Term Memory Network (LSTM). In the decision-level fusion, Weight enumerator and Adaboost techniques were applied to combine the predictions of the CNN and the SVM. The authors have achieved an accuracy of 71% for SVM and 78.56% for the LSTM.

Almeida and Rodrigues [2], developed a system capable of capturing real-time images and alerting the user if there are any signs of stress. The system was divided into several modules, such as:

1. Real-time image capture, via computer camera, sending the images to the next module.
2. Determining the user's face position by the Haar-like feature. The image is further readjusted and normalize.

3. After properly training the classification model, the model will be able to classify each face and return a list of seven classification probabilities, one for each facial expression.
4. The facial expression most likely to be embedded in this module determines whether or not the person is under stress.

The authors have used a CNN previously trained (transfer learning), that was applied using two different techniques: Global Average Pooling (GAP) and Convolution Layer. The classification took into account seven emotions (as in [15]). The authors used two multi-classification models, using them to predict facial expressions and binary classification to classify images with stress/non-stress. They achieved an accuracy of 92% in the best model (VGG16 [27]).

In [5], Bartlett et al. developed a real-time face recognition system that can recognize faces in a video sequence and encode each frame with one of the seven corresponding emotions. All the recognized faces were converted into images of the same size, using Gabor energy filters and were later analyzed by Adaboost, which encodes facial expressions in 7 dimensions (corresponding to 7 emotions). The system was trained and tested with an SVM in the DFAT-504 dataset [13], which is constituted by 100 university students between the ages of 18 and 30: 65% were female, 15% African American, and 3% Asian or Latino. To validate their approach the authors combined a SVM with Adaboost and named it as Adasvm, which produced an accuracy of 93.3%. By themselves, Adaboost achieved an accuracy of 90.1% and SVM 89%, so both were used simultaneously.

Considering situations where the face is partially hidden, mainly the eyes, Cheng et al. [6] used the following approach:

1. Images containing faces are segmented from human images of the same size;
2. The result obtained in the first step is then used to normalize and transform the images into figures of Gabor magnitude through multi-scale and multi-orientation of the Gabor filters. Through these filters the low-level image characteristics of facial figures are reinforced, such as the edges, peaks, contours of crests, eyes, nose and mouth, which are considered to be the main components of the face;
3. The Gabor characteristics are extracted to form a 2D matrix, and the sampling completed downwards to slightly reduce the dimension;
4. The samples are divided into mini-batches and the weights are updated to speed up the pre-training of each Restricted Boltzmann Machine (RBM), which is a bipartite structure with a visible layer and a hidden layer;
5. According to the dimension of the features, set the size of each layer from three layers network (in general);
6. Generate weights and adjust them by fine-tuning;
7. The deep structure training process is divided into two stages: pre-tuning, which treats the data labelled as unmarked for unsupervised training to provide each weight from the lower layer to the top, and fine-tuning which is a simple process of gradient descent under supervision;
8. The test is carried out in number 6 until convergence.

In addition to the Gabor filter, the authors also used other methods that use this filter to compare with the proposed method: Local Gabor Binary Pattern Histogram Sequence (LGBHPS), modified LGBPHS, with the K-Nearest Neighbors (K-NN) Classification method respectively. The authors used the Japanese Female Facial Expression (JAFFE) dataset [9], which consists of 213 images of 10 different individuals with seven different facial expressions: happiness, anger, sadness, fear, surprise, disgust and neutral. Considering that this dataset is not available with natural partial occlusion facial images, the authors simulated the occlusion by overlay graphic masks in the images of this dataset.

The resulting images are divided into 2 parts: 143 images for training and 70 images for testing (images containing occlusion of the eyes, mouth, upper and lower parts of the face). They obtained an accuracy of 85.71% with no occlusion, 82.86% with occlusion of the eyes, 77.14% with oclusion of the upper part of the face and 82.86% with oclusion of the lower part of the face.

Houshmand and Mefraz [12] focused essentially on facial expression identification in the presence of a severe occlusion while the subject is utilizing a head-mounted display in a Virtual Reality (VR) environment. Since display measurements are known, the authors were able to replicate occlusion caused by these headsets using face detection applied to grayscale images using a modification to the conventional Histogram of Oriented Gradients (HOG) and Linear SVM-based approach for object detection. The authors estimated 68 reference coordinates that map the face anatomy on the iBUG 300-W dataset [23]. Because the dataset contains images with varying sizes, the distance between the two temporal bones of the temporal landmarks was used as the reference length, and the polygonal occlusion patch was generated using the midpoint of the line that passes through the center points of the eye as the VR headset's central coordinate.

The authors have evaluated the effectiveness of two different architectures: VGG and ResNet. They chose three datasets that cover different scales of face images and contain images with momentary occlusions to test these two architectures, namely FER+ [4], RAF-DB and Affectnet. The authors have defined a rescale factor of 224 * 224 at the input of the CNNs, which meant that all images used for training, testing, and validation were rescaled for the same measurements and normalized using min-max normalization. The training procedure was carried out by optimizing the multinomial logistic regression objective, which employs mini-batch gradient descent based on momentum back propagation. To regularize the train phase in the ResNet, a max-norm kernel restriction was added. The best result obtained by the authors was 79.98% of accuracy in the ResNet model with transfer learning in the FER+ dataset.

Cornejo et al. [22] have structured their approach in 5-steps. First, preprocessing on all images. This includes the automatic detection of the facial fiducial point, and then the coordinates of the eyes are extracted, the image is rotated, and the image is aligned. Still at this stage, facial expression regions are cut through an appropriate bounding rectangle, RGB images are converted to grayscale, and then randomly generated rectangles are applied over various

regions of the face, such as the left lower eye, right eye, both eyes, right lower side, or lower side. Next, occluded facial expression is reconstructed with the Dual Algorithm based on the principles of the Robust Principal Component Analysis (RPCA) where the Contrast-Limit Adaptive Histogram Equalization (CLAHE) is subsequently applied to reconstructed facial regions, to increase image contrast levels (Fig. 1). Third, a set of facial expression characteristics were extracted through 3 strategies: Weber Local Descriptor (WLD) - applied over the entire facial expression image for extraction of textural features, Local Binary Patterns (LBP) - applied over the entire image to extract histograms of LBP and HOG - applied to the entire image, to extract the HOG features. Fourth, reduction of dimensionality of the characteristics extracted in the previous step, and the resulting descriptor is transferred in a lower dimensional space through PCA and LDA, applied sequentially. Finally, occluded facial expressions are recognized through K-NN and SVM classifiers.

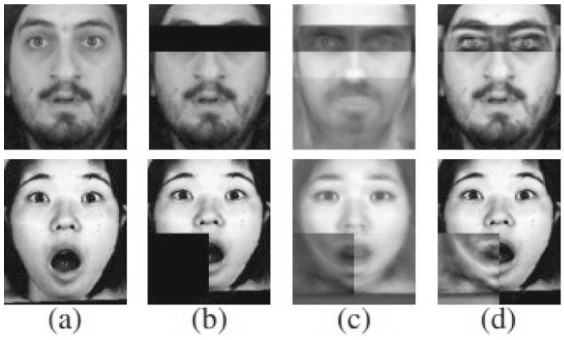


Fig. 1. (a) Cropped images without occlusions from MUG dataset [1] and Jaffe dataset [9]; (b) faces with occlusive areas; (c) reconstructed faces via RPCA; (d) filling the occluded facial regions from (c). [22]

The authors have tested the proposed method on three datasets: the Cohn-Kanade (CK+) [18], JAFFE [9] and Facial MUG Expression [1] and obtained results of 91.01% accuracy in CK+ dataset with K-NN, 92.86% in JAFFE dataset with SVM and 90.1% in MUG with K-NN, in occluded images.

The work described in this paper builds on these, to study the impact of occlusion caused by VR goggles in the identification of facial expressions.

3 Methodology

According to the previous section, the followed approach in this work is based on an CNN ensemble, composed of a ResNet-18 and a VGG19 (Fig. 2) [32]. The weights were initialized with the imagenet weights for both ResNet-18 and VGG19, with xavier for the final fully connected layer.

The same architecture was used in both the scenarios: without occlusion and with occlusion. The chosen dataset was FER2013, composed of 28709 48×48

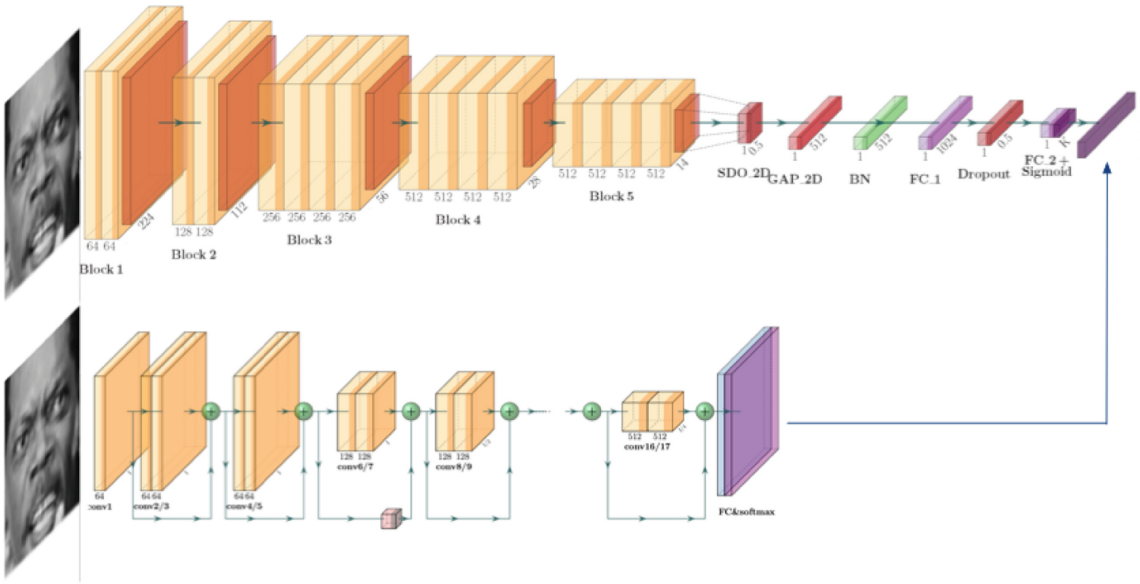


Fig. 2. Ensemble between ResNet-18 and VGG19, with a Fully Connected Layer as output.

grayscale images for training and 3589 for testing, with seven categories (0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, 6 = Neutral). Since the ResNet-18 and VGG19 expect 224×224 RGB images in the input, it was necessary to scale the images up and replicate a single channel image to the remaining two (Fig. 3).

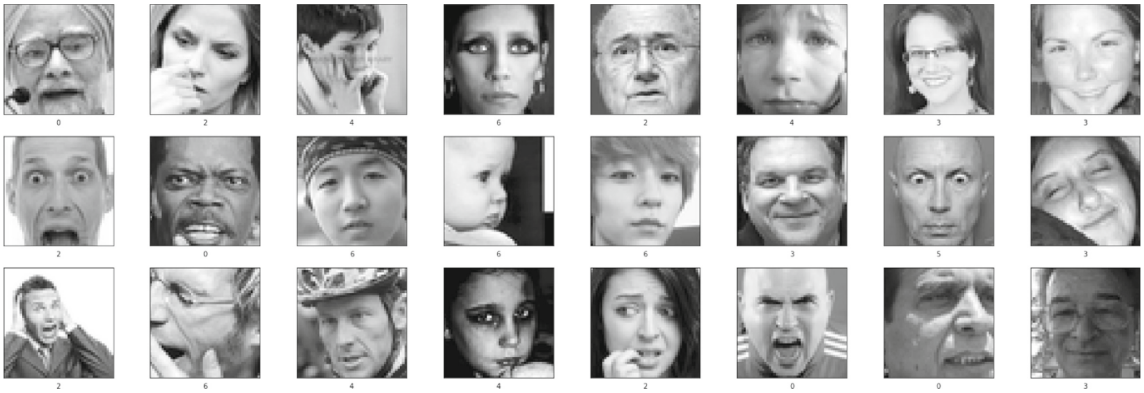


Fig. 3. Sample of the FER2013 dataset.

To simulate the presence of VR goggles, it was necessary to hide the upper part of the face, namely, the eyes. Considering that the face can assume different tilt and yaw positions, the relative position of the eyes change, so the algorithm to calculate the goggles position should contemplate this. For that, it was necessary to obtain location of the face as well as facial landmarks, composed of the position of the eyes, nose and mouth. For that, Multi-task Cascade Convolutional Neural

Networks (MTCNN) was used [31,33]. It consists essentially of 3 parts: i) a network of proposals (P-NET) that foresees potential face positions and their bounding boxes. This process results in a large number of facial detections, many of which are false; ii) a refined network (R-Net) which makes use of the result from step i), thereby refining the result to eliminate most false detection and limiting aggregates; and iii) a network similar to the one used in step ii), called O-Net, further refines the forecasts and adds facials forecasts to the implementation of MTCNN.

With the position of the eyes, the algorithm starts by calculating the middle point between the eyes, and the distance between them (Algorithm 1). It continues by estimating the width and height of the goggles as 20% larger than the distance between the eyes and 150% the distance between the eye line and the nose. The tilt angle is also calculated and, with these, the rectangle is drawn in gray on top of the *sample* image. When the landmarks is empty, meaning that the facial features could not be found, the goggles are not drawn (Fig. 4).

Algorithm 1. Occlusion algorithm.

procedure MAKEGOGGLES(sample, landmarks)
  $left\_eye\_x, left\_eye\_y \leftarrow landmarks[0][0], landmarks[0][5]$
  $right\_eye\_x, right\_eye\_y \leftarrow landmarks[0][1], landmarks[0][6]$
  $nose\_x, nose\_y \leftarrow landmarks[0][2], landmarks[0][7]$
  $middle\_x, middle\_y \leftarrow \frac{right\_eye\_x + left\_eye\_x}{2}, \frac{right\_eye\_y + left\_eye\_y}{2}$
  $googles\_width = 2.2 * \sqrt{(right_eye_y - left_eye_y)^2 + (right_eye_x - left_eye_x)^2}$
  $googles\_height = 1.5 * \sqrt{(middle_eyes_y - nose_y)^2 + (middle_eyes_x - nose_x)^2}$
  $rectangle = (0, 0, googles\_width, googles\_height)$
  $middle\_rectangle\_x, middle\_rectangle\_y = \frac{googles_width}{2}, \frac{googles_height}{2}$
  $angle = \frac{right_eye_y - left_eye_y}{right_eye_x - left_eye_x} * \frac{180}{\pi}$
  $rectangle = rectangle.rotate(-angle, (middle_rectangle_x, middle_rectangle_y))$
  $final_size = rectangle.size$
  $sample.paste(rectangle, \frac{middle_eyes_x - final_size[0]}{2}, \frac{middle_eyes_y - final_size[1]}{2})$

The accuracy of the classification was measured through the confusion matrix and accuracy of each class.

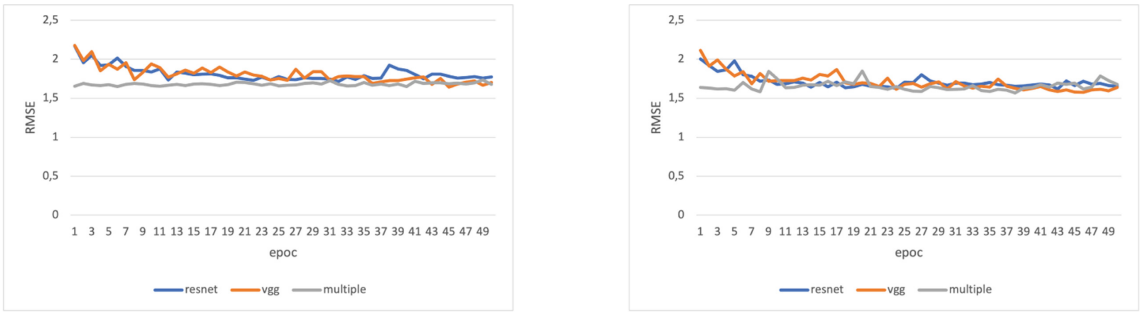
4 Results

Two identical versions of the classification network were trained for 50 epocs with the mini batch of 64 on a AMD Ryzen Threadripper 3970X 32-Core Processor with an NVIDIA GeForce RTX 3090 with 64 GB RAM. The same training and validation datasets were used in both situations, although on the second all the examples were changed to introduce an occlusion over the eyes.

The training process happened in three steps: i) training of the ResNet18; ii) training of the VGG19; iii) training of the full assembly. Both CNNs were

Fig. 4. Sample of the FER2013 dataset with the occlusion algorithm.

initialized with the pretrained ImageNet-1K weights, although all the parameters were allowed to change (Fig. 5). There is a slight better result with the no occlusion dataset, as expected.



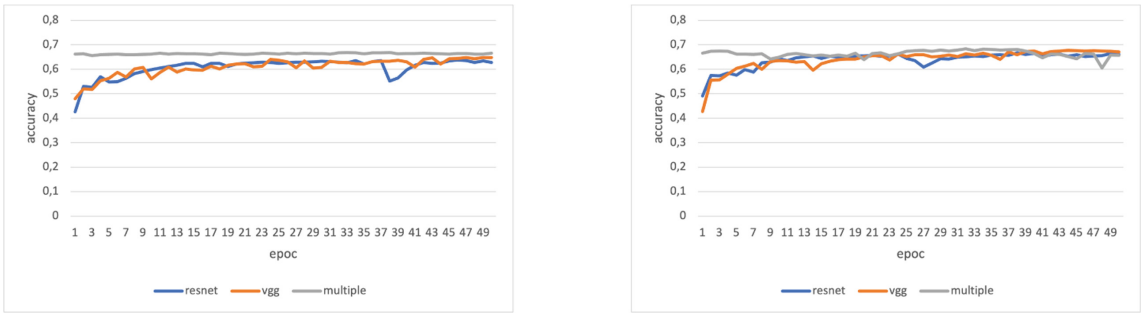
(a) Occlusion (b) No occlusion

Fig. 5. Evolution of RMSE during training

The accuracy also improves with the epochs (Fig. 6). In the occlusion dataset (Fig. 6a) the highest accuracy is obtained with the multiple model (consisting on the combination of both models). However, in the no occlusion situation (Fig. 6b) it is the lowest. Looking at the progress during the epochs, it seems that the increase in the number of epochs would achieve better accuracy.

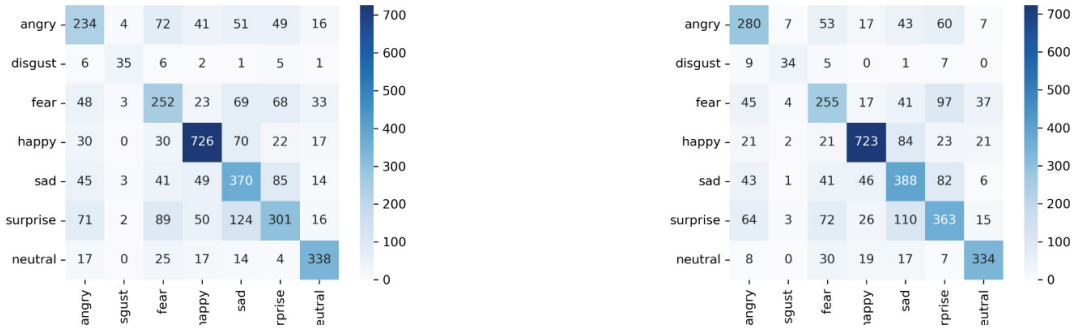
To illustrate the classification capacity of the three models in predicting the seven facial expressions in the FER-2013 dataset, confusion matrices were built comparing the situation between occlusion and no occlusion datasets (Figs. 7, 8 and 9).

The classes are not balanced so, intra-class normalization was also performed (Table 1). The impact of occlusion of the eyes is, apparently, marginal. The highest accuracy in both situations (occlusions and no occlusion) is in the class 'neutral', immediately followed by 'happy' in the occlusion and 'sad'/'surprise' for the no occlusion. The lowest accuracy was in the classes 'disgust' and 'fear',



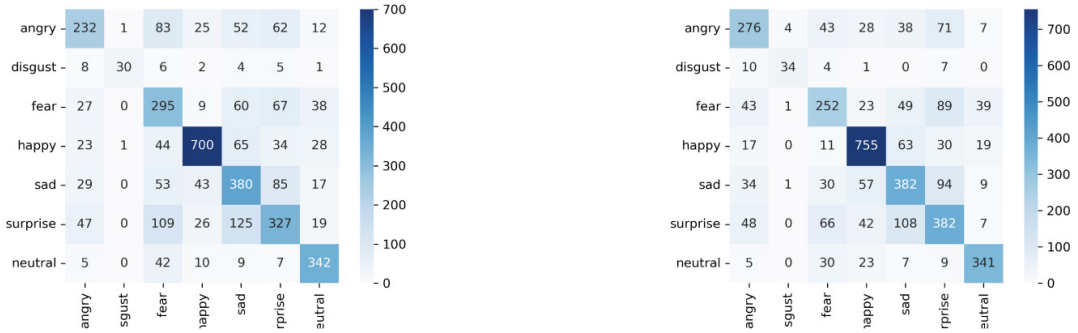
(a) Occlusion (b) No occlusion

Fig. 6. Evolution of the accuracy during training



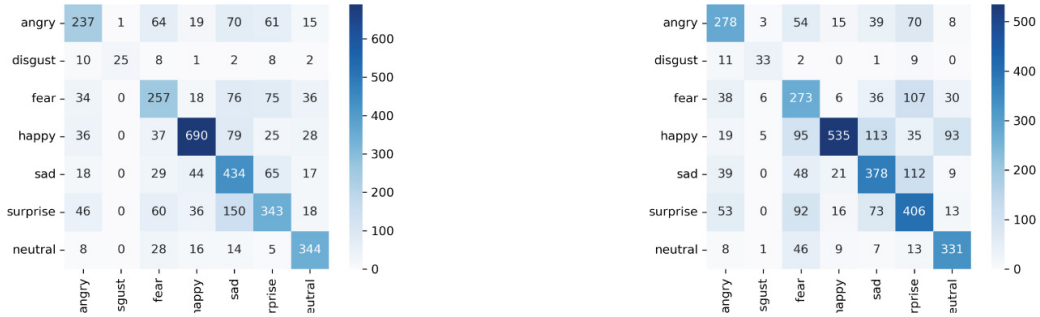
(a) Occlusion (b) No occlusion

Fig. 7. Confusion matrix for the ResNet18



(a) Occlusion (b) No occlusion

Fig. 8. Confusion matrix for the VGG19



(a) Occlusion (b) No occlusion

Fig. 9. Confusion matrix for the combined model

respectively. According to the previous confusion matrices, the 'disgust' class is often misclassified as 'angry' or 'surprise'.

Table 1. Accuracy per class

	(a) Occlusion			(b) No occlusion		
Class	**ResNet18**	**VGG19**	**Combined**	**ResNet18**	**VGG19**	**Combined**
angry	0,501	0,497	0,508	0,6	0,591	0,595
disgust	0,625	0,536	0,446	0,607	0,607	0,589
fear	0,508	0,595	0,518	0,514	0,508	0,55
happy	0,811	0,782	0,771	0,808	0,844	0,598
sad	0,61	0,626	0,715	0,639	0,629	0,623
surprise	0,461	0,501	0,525	0,556	0,585	0,623
neutral	0,814	0,824	0,829	0,805	0,822	0,798
F1-score	0,627	0,645	0,649	0,663	0,673	0,628

The overall accuracy of the combined model was 61.6% for the occlusion and 62.5% for the no occlusion, which is far from the state of the art result. Nevertheless, the purpose of this work was to assess the impact of occlusion and the results confirm that most of the classification is performed with the mouth and chin.

5 Conclusions

In this paper, a residual network (ResNet), a VGG and the combination of both models, were used to predict facial expressions in real-time, occluding part of the face (eyes and the upper part of the nose). The model achieved an accuracy of 64.9% for the occlusion dataset and 62.8% no occlusion, using the FER-2013 dataset. As this dataset only contains frontal images of the face, the correct evaluation of facial expression is not performed if the individual's face is not located in a position other than the frontal, and it also does not perform the

correct classification of people wearing glasses or other accessories. Similarly, it was also observed that the MTCNN does not correctly locate the face in real-time, that is, any other object is often identified as a face. It was also found that with a small movement made in real-time (such as, for example, the upward movement of the muscles of the cheek and the sides or edges of the lips to form a smile that indicates, to the human eye, that the person is happy), the model is not very robust in its classification. That means that it should classify this movement as 'Happy', instead of classifying it as 'Neutral', 'Angry' and 'Sad', only in a small fraction of the time, it is classified correctly as 'Happy'. For the above-mentioned situations, it is proposed to study other datasets as well as other classifiers that allow evaluating facial expressions, the adjustment of the MTCNN to enable correct face detection and the addition of more MTCNN tools to be also able to locate the eyes and lips.

Acknowledgment. This work is funded by the European Regional Development Fund (ERDF) through the Regional Operational Program North 2020, within the scope of Project GreenHealth - Digital strategies in biological assets to improve well-being and promote green health, Norte-01-0145-FEDER-000042. This work has been supported by FCT - Fundação para a Ciência e Tecnologia within the Project Scope: UIDB/05757/2020.

References

1. Aifanti, N., Papachristou, C., Delopoulos, A.: The MUG facial expression database. In: 11th International Workshop on Image Analysis for Multimedia Interactive Services WIAMIS 10, pp. 1–4. IEEE (2010)
2. Almeida, J., Rodrigues, F.: Facial expression recognition system for stress detection with deep learning. In: Proceedings of the 23rd International Conference on Enterprise Information Systems, pp. 256–263. SCITEPRESS - Science and Technology Publications, Online Streaming, – Select a Country – (2021). https://doi.org/10.5220/0010474202560263. https://www.scitepress.org/DigitalLibrary/Link.aspx?doi=10.5220/0010474202560263
3. Baltrusaitis, T., Robinson, P., Morency, L.P.: OpenFace: an open source facial behavior analysis toolkit. In: 2016 IEEE Winter Conference on Applications of Computer Vision (WACV), Lake Placid, NY, USA, pp. 1–10. IEEE, March 2016. https://doi.org/10.1109/WACV.2016.7477553. https://ieeexplore.ieee.org/document/7477553/
4. Barsoum, E., Zhang, C., Ferrer, C.C., Zhang, Z.: Training deep networks for facial expression recognition with crowd-sourced label distribution. In: Proceedings of the 18th ACM International Conference on Multimodal Interaction, Tokyo, Japan, pp. 279–283. ACM, October 2016. https://doi.org/10.1145/2993148.2993165. https://dl.acm.org/doi/10.1145/2993148.2993165
5. Bartlett, M., Littlewort, G., Lainscsek, C., Fasel, I., Movellan, J.: Machine learning methods for fully automatic recognition of facial expressions and facial actions. In: 2004 IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No. 04CH37583), The Hague, Netherlands, vol. 1, pp. 592–597. IEEE (2004). https://doi.org/10.1109/ICSMC.2004.1398364. https://ieeexplore.ieee.org/document/1398364/

6. Cheng, Y., Jiang, B., Jia, K.: A deep structure for facial expression recognition under partial occlusion. In: 2014 Tenth International Conference on Intelligent Information Hiding and Multimedia Signal Processing, pp. 211–214 (2014). https://doi.org/10.1109/IIH-MSP.2014.59
7. Devries, T., Biswaranjan, K., Taylor, G.W.: Multi-task Learning of Facial Landmarks and Expression. In: 2014 Canadian Conference on Computer and Robot Vision, Montreal, QC, Canada, pp. 98–103. IEEE, May 2014. https://doi.org/10.1109/CRV.2014.21. https://ieeexplore.ieee.org/document/6816830/
8. Donaldson, M.: Plutchik's wheel of emotions-2017. Update (2017)
9. Cheng, F., Yu, J., Xiong, H.: Facial expression recognition in JAFFE dataset based on gaussian process classification. IEEE Trans. Neural Netw. **21**(10), 1685–1690 (2010)
10. Goodfellow, I.J., et al.: Challenges in representation learning: a report on three machine learning contests. Neural Netw. **64**, 59–63 (2015)
11. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-PIE. Image Vis. Comput. **28**(5), 807–813 (2010)
12. Houshmand, B., Mefraz Khan, N.: Facial expression recognition under partial occlusion from virtual reality headsets based on transfer learning. In: 2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM), New Delhi, India, pp. 70–75. IEEE, September 2020. https://doi.org/10.1109/BigMM50055.2020.00020. https://ieeexplore.ieee.org/document/9232653/
13. Kanade, T., Cohn, J., Tian, Y.: Comprehensive database for facial expression analysis. In: Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580), pp. 46–53. IEEE Comput. Soc, Grenoble, France (2000). https://doi.org/10.1109/AFGR.2000.840611. https://ieeexplore.ieee.org/document/840611/
14. Li, R., et al.: MindLink-Eumpy: an open-source Python toolbox for multimodal emotion recognition. Front. Hum. Neurosci. **15**, 621493 (2021)
15. Loizou, C.P.: An automated integrated speech and face imageanalysis system for the identification of human emotions. Speech Commun. **130**, 15–26 (2021)
16. Lopes, J.C., Lopes, R.P.: A review of dynamic difficulty adjustment methods for serious games. In: Pereira, A.I., et al. (eds.) OL2A 2022, CCIS 1754, pp. xx–yy (2022)
17. Lopes, R.P., et al.: Digital technologies for innovative mental health rehabilitation. Electronics **10**(18) (2021). https://doi.org/10.3390/electronics10182260. https://www.mdpi.com/2079-9292/10/18/2260
18. Lucey, P., Cohn, J.F., Kanade, T., Saragih, J., Ambadar, Z., Matthews, I.: The extended Cohn-Kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops, pp. 94–101. IEEE (2010)
19. Mehrabian, A.: Communication without words. In: Mortensen, C.D. (ed.) Communication Theory, Routledge, 2 edn., pp. 193–200, September 2017. https://doi.org/10.4324/9781315080918-15. https://www.taylorfrancis.com/books/9781351527538/chapters/10.4324/9781315080918-15
20. Poria, S., Majumder, N., Mihalcea, R., Hovy, E.: Emotion recognition in conversation: research challenges, datasets, and recent advances. IEEE Access **7**, 100943–100953 (2019)
21. Prodger, P.: Darwin's Camera: Art and Photography in the Theory of Evolution. Oxford University Press, Oxford (2009)

22. Ramirez Cornejo, J.Y., Pedrini, H.: Emotion recognition from occluded facial expressions using weber local descriptor. In: 2018 25th International Conference on Systems, Signals and Image Processing (IWSSIP), Maribor, Slovenia, pp. 1–5. IEEE, June 2018. https://doi.org/10.1109/IWSSIP.2018.8439631. https://ieeexplore.ieee.org/document/8439631/
23. Sagonas, C., Tzimiropoulos, G., Zafeiriou, S., Pantic, M.: 300 faces in-the-wild challenge: the first facial landmark localization challenge. In: 2013 IEEE International Conference on Computer Vision Workshops, Sydney, Australia, pp. 397–403. IEEE, December 2013. https://doi.org/10.1109/ICCVW.2013.59. https://ieeexplore.ieee.org/document/6755925/
24. Saurav, S., Saini, A., Saini, R., Singh, S.: Deep learning inspired intelligent embedded system for haptic rendering of facial emotions to the blind. Neural Comput. Appl. **34**(6), 4595–4623 (2022). https://doi.org/10.1007/s00521-021-06613-3. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85117830803&doi=10.1007%2fs00521-021-06613-3&partnerID=40&md5=ff31d4bd7bc4190b4483b813b2837a34, publisher: Springer Science and Business Media Deutschland GmbH
25. Singh, S., Gupta, A., Pavithr, R.S.: Automatic classroom monitoring system using facial expression recognition. In: Sanyal, G., Travieso-González, C.M., Awasthi, S., Pinto, C.M.A., Purushothama, B.R. (eds.) International Conference on Artificial Intelligence and Sustainable Engineering. LNEE, vol. 836, pp. 151–165. Springer, Singapore (2022). https://doi.org/10.1007/978-981-16-8542-2_12. https://www.scopus.com/inward/record.uri?eid=2-s2.0-85130262593&doi=10.1007%2f978-981-16-8542-2_12&partnerID=40&md5=3d727b02a4b6cbca64032f67f9156366. ISBN: 9789811685415
26. Susskind, J.M., Anderson, A.K., Hinton, G.E.: The Toronto face database. Department of Computer Science, University of Toronto, Toronto, ON, Canada, Technical report 3 (2010)
27. Tammina, S.: Transfer learning using VGG-16 with deep convolutional neural network for classifying images. Int. J. Sci. Res. Publ. (IJSRP) **9**(10), 9420 (2019). https://doi.org/10.29322/IJSRP.9.10.2019.p9420. https://www.ijsrp.org/research-paper-1019.php?rp=P949194
28. Tang, Y.: Deep learning using linear support vector machines, February 2015. arXiv:1306.0239 [cs, stat]
29. Viana, I.: Comunicação não verbal e expressões faciais das emoções básicas. Revista de Letras **13**(II), 165–181 (2014)
30. Wood, E., Baltruaitis, T., Zhang, X., Sugano, Y., Robinson, P., Bulling, A.: Rendering of eyes for eye-shape registration and gaze estimation. In: 2015 IEEE International Conference on Computer Vision (ICCV), Santiago, Chile, pp. 3756–3764. IEEE, December 2015. https://doi.org/10.1109/ICCV.2015.428. https://ieeexplore.ieee.org/document/7410785/
31. Xiang, J., Zhu, G.: Joint face detection and facial expression recognition with MTCNN. In: 2017 4th International Conference on Information Science and Control Engineering (ICISCE), pp. 424–427 (2017). https://doi.org/10.1109/ICISCE.2017.95
32. Yu, Z., Zhang, C.: Image based static facial expression recognition with multiple deep network learning. In: Proceedings of the 2015 ACM on International Conference on Multimodal Interaction, Seattle, Washington, USA, pp. 435–442. ACM, November 2015. https://doi.org/10.1145/2818346.2830595. https://dl.acm.org/doi/10.1145/2818346.2830595

33. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
34. Zhang, X., Sugano, Y., Fritz, M., Bulling, A.: Appearance-based gaze estimation in the wild. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, pp. 4511–4520. IEEE, June 2015. https://doi.org/10.1109/CVPR.2015.7299081. https://ieeexplore.ieee.org/document/7299081/
35. Zhao, X., Zhang, S.: A review on facial expression recognition: feature extraction and classification. IETE Techn. Rev. **33**(5), 505–517 (2016)