ISSN 2669-0330 (online)
ĮŽVALGOS
2024 Nr. 2, p. 85-97

REAL-TIME EMOTION RECOGNITION USING DEEP NEURAL NETWORKS AND APPLICATION FOR CHILDREN'S RELAXATION SYSTEM

Eugenijus Mačerauskas[1], Maksims Žigunovs[2]
[1]*Utenos kolegija,*
7 Maironio str., Utena, Lithuania
[2] *Ryga Technical University Liepaja Academy,*
14 Liela iela, Liepaja, Latvia

Annotation

The task of recognizing emotions from facial images obtained from a video stream is considered. The approach to the solution is based on the use of deep neural networks. A dataset used for network training, its characteristics, and the distribution of data by emotion classes are presented. Two convolutional neural network models are described: a classical convolutional neural network built for this task and a convolutional neural network improved by regularization mechanisms. Based on the obtained network training results, a comparative analysis of classification accuracy is carried out. The process of recognizing emotions on arbitrary data not related to the dataset in question is described. The emotion recognition algorithm was realized on a hardware platform and produced positive outcomes. An example of applying an emotion recognition algorithm in a real hardware system is presented.

Keywords: emotion recognition, deep machine learning, highly accurate neural networks, highly accurate neural networks.

Introduction

Scientists have been studying emotions and their manifestations for quite a long time. After all, emotions are an inevitable part of any interpersonal communication; they express a person's attitude to the world around them, to the situation around them, and to themselves. At the same time, the need to identify human emotions has recently increased even more. This is primarily due to the expansion of the scope of application of the emotion recognition problem. Currently, this includes driver monitoring, smart city video analytics systems, marketing research, and security systems. In some cases, emotion recognition is relevant when developing relaxation systems or psychologist-assistant software.

Emotions can be expressed in different ways: facial expressions, voice, behavior, and reactions of body systems (Gaind, B. 2019). Of greatest interest among them is the recognition of human emotions by their facial expression. This task is quite popular now for several reasons: such images are easy to obtain, they contain a lot of useful information for recognizing emotions, and it is quite easy to collect a large dataset in the form of facial images (compared to other material for recognition: speech or handwriting samples).

This work is devoted to the task of recognizing emotions from images of a human face. For the full research cycle - the formation of a dataset, the creation, training, and testing of models - Python was used as one of the most popular languages for solving problems in the field of data analysis and machine learning. The developed AI model was applied to the development of a relaxation system for children with emotional disorders.

1. Dataset

The Facial Expression Recognition 2013 (FER2013) dataset, which was presented at the International Conference on Machine Learning 2018 (Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. 2018), was chosen as the dataset for training deep networks. This dataset contains 35,887 images with a resolution of 48x48 pixels, most of which were taken in uncontrolled conditions. The database was created using Google image search tools. Each image is classified as one of seven types of emotions: surprise, fear, happiness, anger, disgust, sadness, and a neutral state or calmness. FER has a large number of variations in the images, including partial occlusion of the face (mostly by hand). The distribution of data across different emotion classes and examples of facial images with an indication of the classes to which they are assigned are presented in Figure 1.

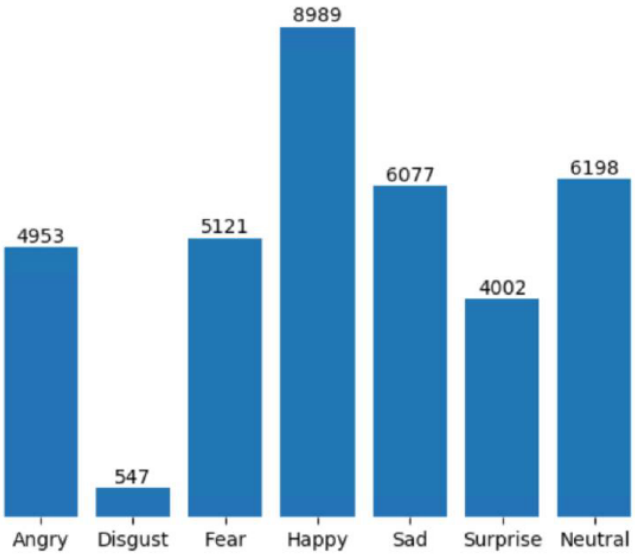


Figure 1. The diagram of data distribution across different emotion classesin FER 2013 dataset.
Source: FER2013 Dataset

Some examples from the database are shown in Figure 2.

Figure 2. An examples of facial images with an indication of the classes in FER 2013 dataset
Source: FER2013 Dataset

The FER 2013 dataset is divided into three parts: a training set, a validation set, and a test set. The first two participate in network training: the training set is used to optimize the model weights, and the validation set provides metrics after each training epoch, 1 which help to assess the quality of the model training. The test set is necessary to compare the recognition accuracy among different models.

2. The Architecture of Neural Networks

For emotion recognition, this work utilizes a Convolutional Neural Network (CNN) architecture. Schematically, a CNN is a sequence of layers. Each layer transforms one activation volume into another using a differentiable function. Three main layers are used to organize a convolutional neural network: convolution, pooling (also known as subsampling or downsampling), and fully connected (FC) layers. Convolution and pooling layers are used to extract feature maps from the original image, while fully connected layers are used for the final classification of the image based on the extracted features.

The size of the network's input layer is 48x48x1, in accordance with the size of the images from the dataset. The output layer of the network is a vector of 7 elements, corresponding to the probabilities of the input image belonging to each of the classes. As a result, the input image is assigned to the class with the maximum probability value.

Two CNN models were built during the study. The first model contains 2 convolutional layers, 2 pooling layers, and 4 fully connected layers. A detailed illustration of the first model: layer dimensions, their parameters, and the activation functions used are presented in Figure 3.

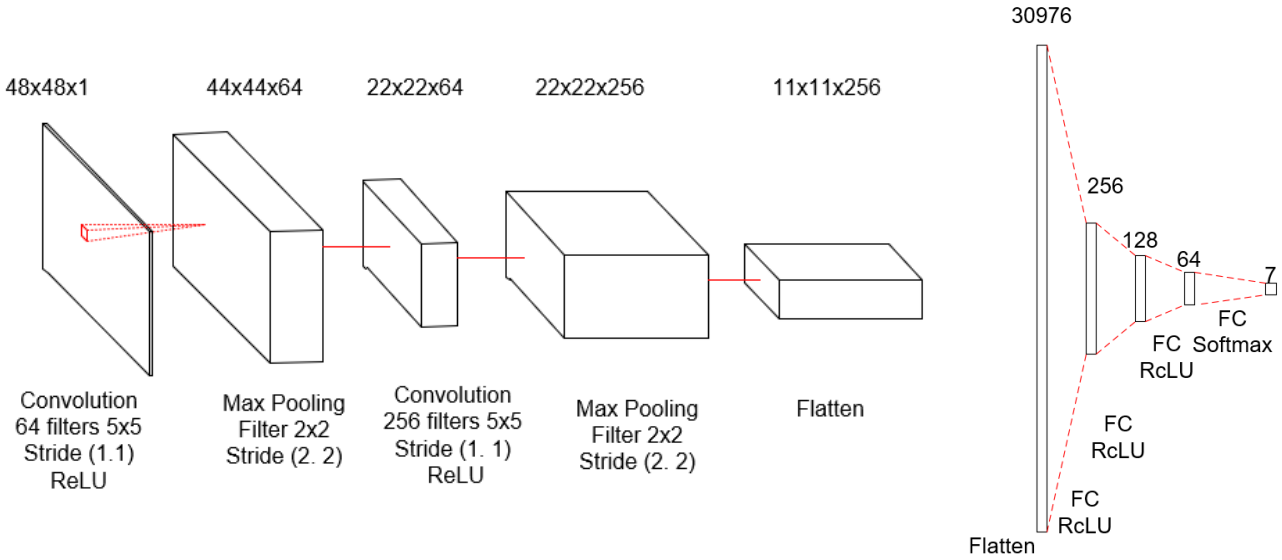


Figure 3. The model of the first Neutral Network.

Source: created by the authors

The second model is a modification of the first model and contains 8 convolutional layers, 4 pooling layers, and 4 fully connected layers, as well as a regularization mechanism (Shukla, V., & Choudhary, S. 2022). An illustration of the second model is presented in Figure 4.

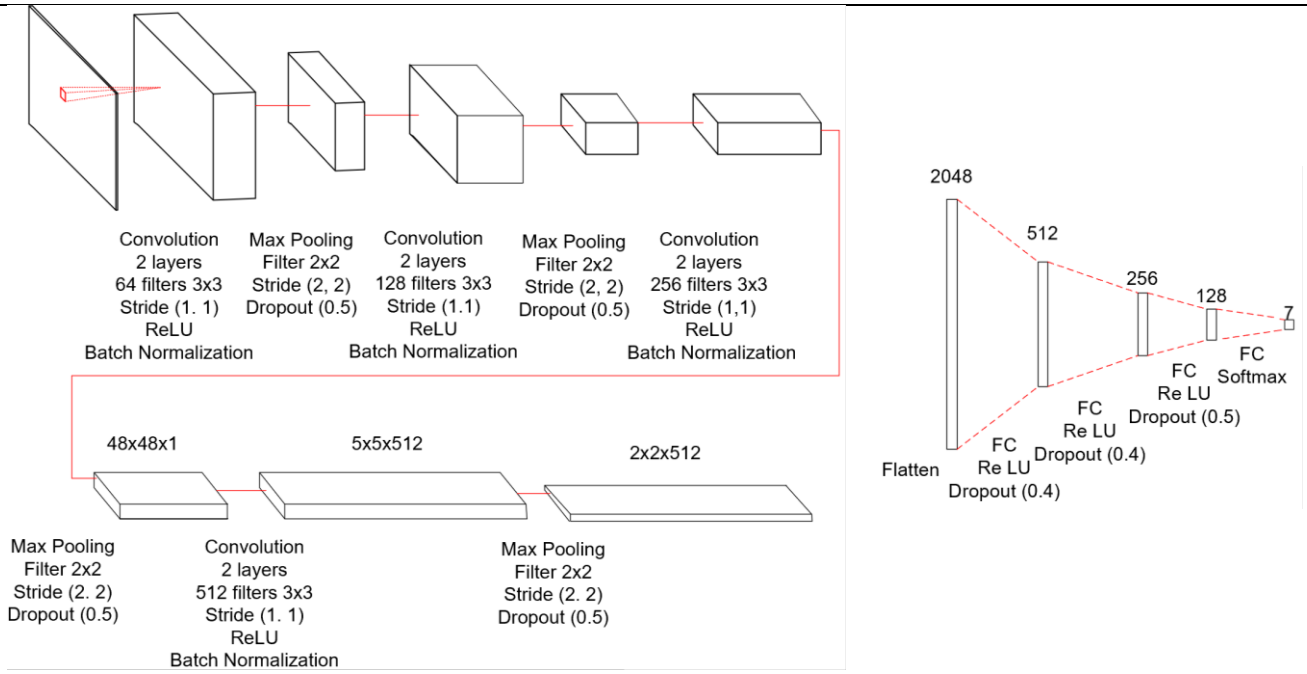


Figure 4. The model of the first Neutral Network.

Source: created by the authors

Compared to the first model, it has a larger number of convolutional layers that have a smaller convolution matrix size, which allows for the extraction of a more detailed feature map. The regularization mechanism helps to avoid a situation called overfitting (Salman, S. 2020). A characteristic feature of overfitting is high recognition accuracy on the training set and relatively low recognition accuracy on the test set. This situation can arise if the data has many features, but the dataset itself contains few examples, or when the model is too complex for the data. To prevent overfitting in the second network model, regularization mechanisms such as Batch Normalization (Ioffe, S. 2020) and Dropout (Sabiri, B., El Asri, B., & Rhanoui, M. 2022) are used. Let's consider the ideas behind these mechanisms.

Typically, some preprocessing of the input data is performed to train a neural network. For example, the FER dataset is normalized so that its data resembles a normal distribution - it has zero mean and unit variance. This processing is done to prevent early saturation of the nonlinear activation functions of the layers and to ensure that all input data are in the same range of values. But the problem arises in the intermediate layers, as the distribution of values that the activation function can have is constantly changing during training. This slows down the learning process because each layer has to learn to adapt to a new distribution at each training step. This problem is known as internal covariate shift.

The essence of the Batch Normalization method is to normalize the input values of the inner layers of the neural network and thus prevent the occurrence of internal covariate shift. During training, the Batch Normalization mechanism performs the following actions.

Calculate the mean $\boldsymbol{\mu_B}$ and variance $\boldsymbol{\sigma^2}$ of the input values of the layer (1):

$$\begin{cases} \mu_B = \frac{1}{m}\sum_{i=1}^{m} x_i ; \\ \sigma_B^2 = \frac{1}{m}\sum_{i=1}^{m} (x_i - \mu_B). \end{cases} \quad (1)$$

where: $\boldsymbol{m}$ - sample size, which is the number of data points in sample, $\boldsymbol{x_i}$ standardized value of the i-th data point in the batch.

The input values $\bar{x}_i$ of the layer are normalized using the previously calculated statistical values (2):

$$\bar{x}_i = \frac{x_i - \mu_B}{\sqrt{\sigma_B^2}} \quad (2)$$

The normalized values are scaled and shifted to avoid changing the representation of the data in the layer (3):

$$y_i = \gamma \bar{x}_i + \beta \quad (3)$$

where: the scaling parameters γ and shift β are adjusted during training along with other network parameters.

The main idea behind the Dropout mechanism is to randomly drop individual neurons (along with their connections) from the neural network during training. Since the dropped neurons no longer contribute to the network training process, this becomes equivalent to training a new neural network. This prevents the neurons from adapting too much to each other. Each layer that uses Dropout has a parameter that determines the probability of a neuron being excluded from the network.

3. The Architecture of Neural Networks

After training, the first network demonstrated an emotion recognition accuracy of 52% on the test dataset. At the same time, the recognition accuracy on the training dataset was 98%. The graph of the accuracy change during the model training process is shown in Figure 5.

The confusion matrix built on the test dataset is shown in Figure 6. In the confusion matrix, rows and columns are labeled with one of the seven emotion classes. The intersection indicates the number of instances assigned to the emotion class denoting the column, but actually belonging to the emotion class denoting the current row. The matrix shows that the recognition of the "happiness" emotion is subject to the least error (23% errors), and the recognition of the "fear" and "anger" emotions is subject to the greatest error (65% errors).

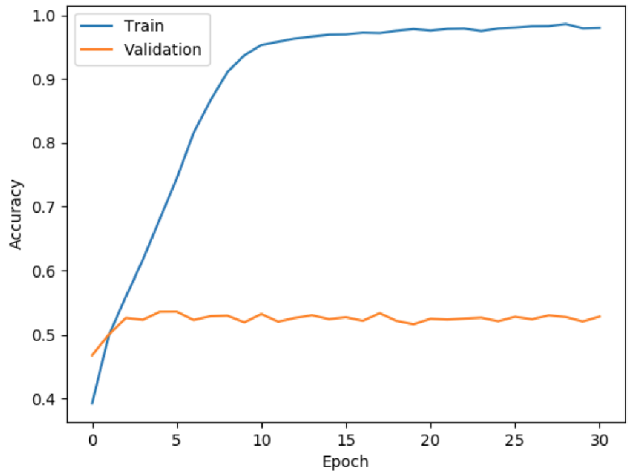


Figure 5. Graph of the change in the model's recognition accuracy depending on the training epoch.

Source: created by the authors

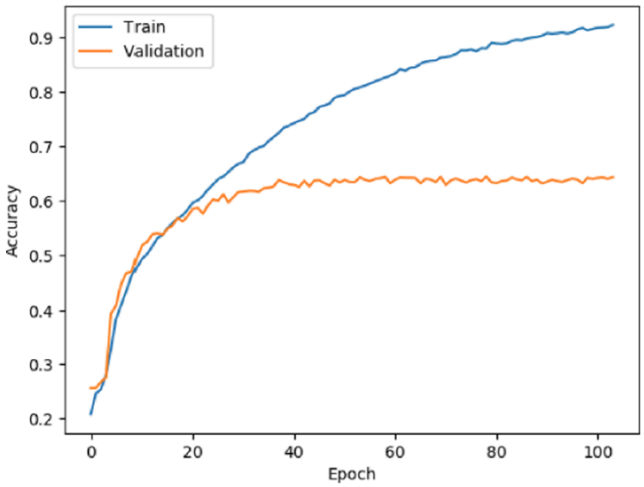


Figure 6. Model's recognition accuracy as a function of the training epoch number

Source: created by the authors

A graph that shows how well a model performs (its accuracy in recognizing something) over time as it learns (each training cycle is called an epoch).

The second convolutional network uses the same dataset as the first network. As a result of training, the network shows a recognition accuracy of 92% on the training dataset. However, on the validation dataset, the accuracy reaches 64% (Fig. 6). The correlation between correct and incorrect recognitions is presented by the confusion matrix in Figure 7.

Analysis of the results allows us to conclude that regularization, together with the addition of new convolutional layers in the model, improved the recognition accuracy by 12%

The high recognition accuracy on the training set and the relatively low accuracy on the test set are signs of network overfitting. As mentioned earlier, the solution to this problem is the regularization mechanism, which is added to the second convolutional network.

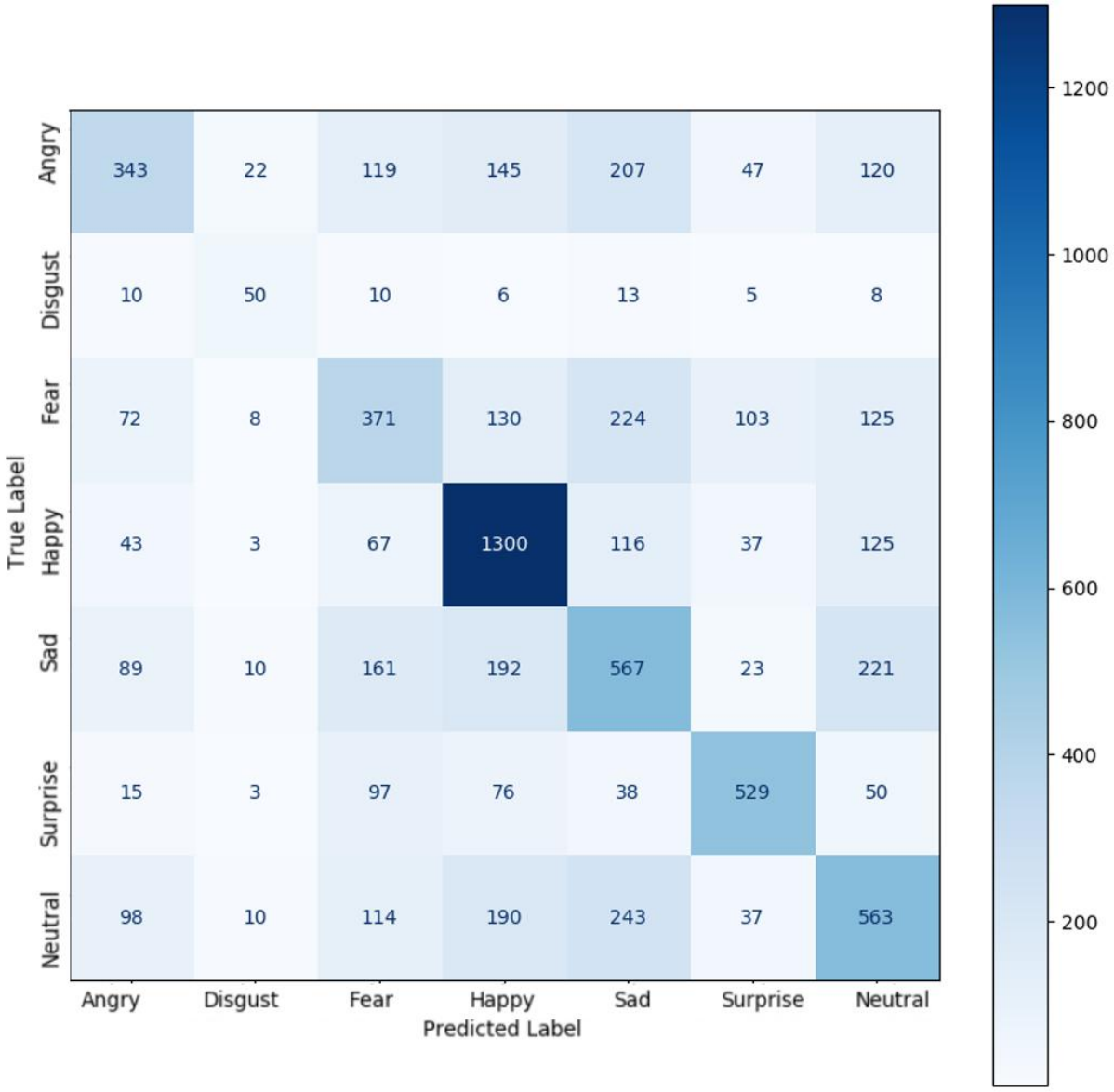


Figure 7. Erors matrix of the pirst network model

Source: created by the authors

4. Testing the Trained Model on Arbitrary Data

To test the trained model on arbitrary data, an auxiliary application was developed that allows classifying emotions on a given image or video. The data source can be either a pre-recorded video or a video coming from a camera in real-time.

The OpenCV library (Culjak, I., et al. (2012). is used for decoding and frame-by-frame video processing. Face detection on a separate frame is performed using the Viola-Jones method (Viola, P. (2001). This method demonstrates high accuracy in finding faces in an image along with fast operation speed. There are also alternative face detection methods based on convolutional neural networks, but they require more resources to process the image (Deva Priya W. (2022), making the Viola-Jones method a more acceptable option for classifying emotions in real-time with a high frame rate.

After performing face detection using the Viola-Jones method, all faces found in the frame undergo a series of transformations to improve the accuracy of further classification:

Next, the transformed set of faces is fed to the input of the classifier - the trained model. After the classification is complete, each face image will be assigned a corresponding emotion class. The final stage of frame processing is the visualization of the obtained classes - each detected face in the frame is marked with a colored frame and labeled with the name of the assigned emotion.

The frame processing process is schematically shown in Figure 8.

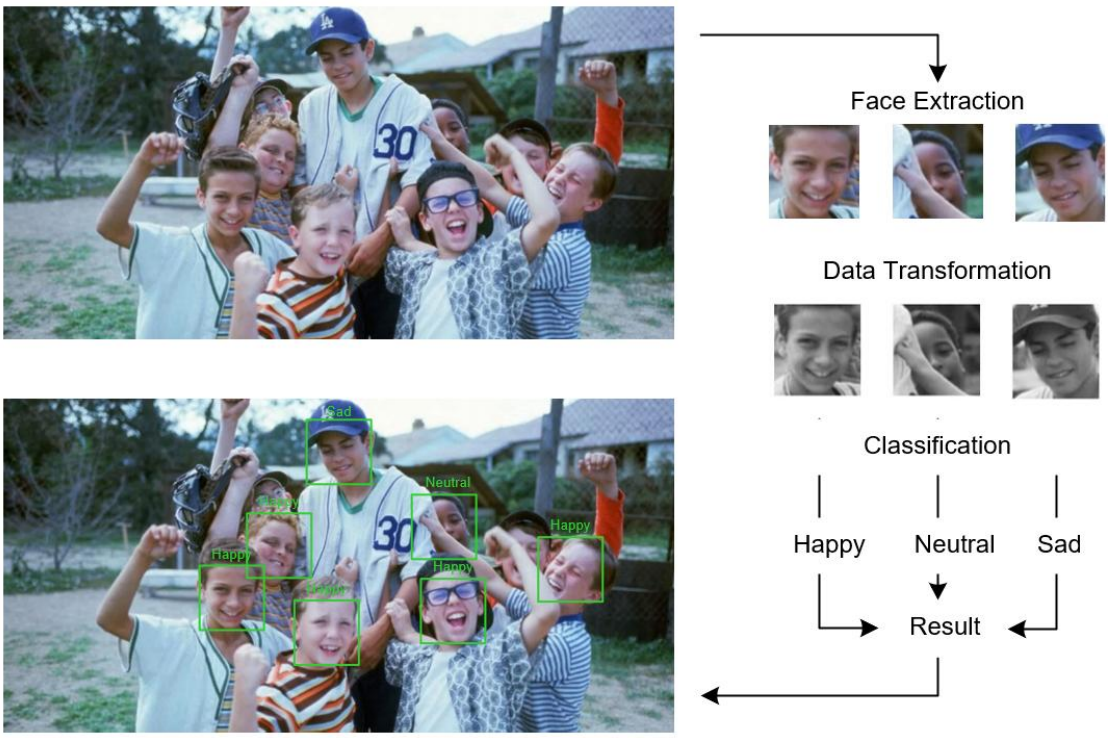


Figure 8. Emotion detection in video stream.

Source: created by the authors

As a result of the network design and subsequent training, the best achieved accuracy of emotion classification from facial images was 64%. The constructed error matrix demonstrates that the achieved classification accuracy is primarily due to the uneven distribution of data across classes in the original dataset. For example, the number of images assigned to the "disgust" class is 16 times less than the number of images assigned to the "happiness" class.

Testing the model on arbitrary data not belonging to the FER dataset allowed for a qualitative assessment of the emotion recognition accuracy. It was revealed that due to the low resolution of the input image, an error occurs in the model's recognition.

Further research will be aimed at both improving the dataset used and developing the current convolutional neural network model.

5. Experimental Hardware Implementation of an Algorithm

Based on the previously described neural network models: Model 1 (according to Figure 3) and Model 2 (according to Figure 4), software was developed in the Python programming language using the OpenCV library. Two identical software packages operated on identical hardware emotion detection systems. The only difference was the neural network portion of the software. The experimental setup is shown in Figure 9.

Each system (Recognition system_1 and Recognition system_2) consisted of a Raspberry Pi 4 (8GB RAM) and a 'Pi Camera Module v2' video camera with a 'Sony IMX219' image sensor. The video resolution was software-limited to 640x480 pixels to reduce the amount of processed data.

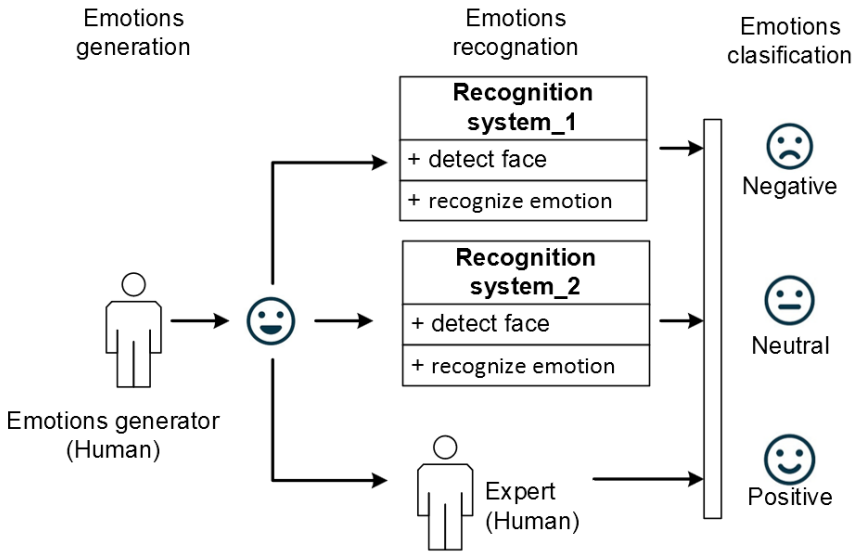


Figure 9. Schematic diagram of the experiment system

Source: created by the authors

Three volunteer students (two females and one male) participated in the experiment, displaying 160 different facial expressions. A human expert, whose prediction accuracy is considered 100%, evaluated the performance of each system. The experiment results are presented in Table 1.

Table 1. **Results of the experiment**

Emotion Category	Human (100%Accuracy)	System 1 (69% Accuracy)	System 2 (94% Accuracy)	Emotion Category
Negative	53	37	50	Negative
Neutral	53	37	50	Neutral
Positive	54	36	50	Positive

Based on the results presented in Table 1, the following conclusions can be drawn.

System 2, with its 94% accuracy, significantly outperformed System 1, which achieved only 69% accuracy. This clearly demonstrates a substantial difference in the neural networks models' ability to accurately classify emotional states.

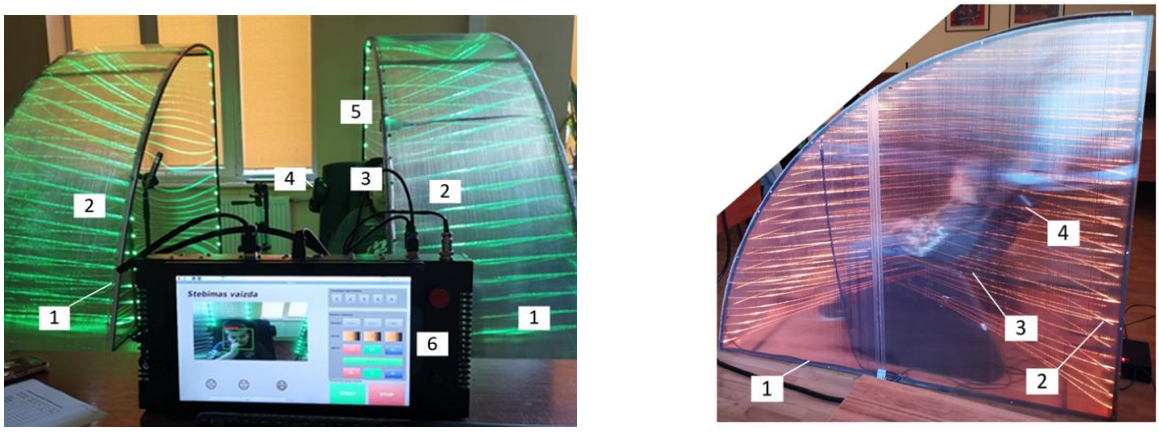
The accuracy metric effectively highlighted the disparity in the models' performance, providing a clear and quantifiable measure of their emotional recognition capabilities.

The high accuracy of Model 2 suggests its potential for practical applications requiring precise emotional state classification, such as sentiment analysis, human-computer interaction, and psychological research.

6. Practical Implementation

The developed model was empirically validated through its integration into a relaxation system designed for children with emotional disorders. The model was implemented within software operating on a Raspberry Pi embedded computing platform, an integral component of the relaxation system. This system is designed to modulate children's emotional states using controlled sensory stimuli, including light, audio output, and tactile vibration, all managed by an emotion recognition algorithm. The external hardware configuration of the relaxation system is illustrated in Figure 10.

The system consists of a vibrating chair (4), which incorporates mechanical vibration actuators and an integrated audio output module. Surrounding the chair is a laterally sliding, semi-transparent enclosure (1), which includes a distributed array of colored light sources (2) along its perimeter. A static video capture device (3) is affixed to one segment of the enclosure, providing facial image data to an artificial intelligence-driven relaxation scenario selection algorithm. System operations are managed through a dedicated computer interface (5) equipped with a touch screen display. Power is supplied by an external power source (6), remotely located to ensure user safety, delivering low-voltage electrical current to minimize potential hazards.



a) b)

Figure 10. Real Prototype Image with the Subject Child:
a) - Frontal Projection, b) - Lateral Projection

1 - Semi-transparent enclosure, 2 - Lighting system, 3 - Relaxation chair, 4 - Sound system, 5 - Video camera, 6 - Control computer.

Source: created by the authors

The system's operational principle is predicated on inducing relaxation in children exhibiting emotional lability, such as agitation or anxiety. This is achieved through a multi-sensory stimulation paradigm, wherein the subject occupies a vibroacoustic chair. The chair provides synchronized

vibrotactile stimulation, coordinated with auditory output from an integrated sound system and chromotherapeutic illumination from a dynamic light array. This combined sensory input facilitates a rapid return to emotional homeostasis, significantly expedited compared to conventional methods. The functional interrelation of the relaxation system's constituent components is illustrated in Figure 11.

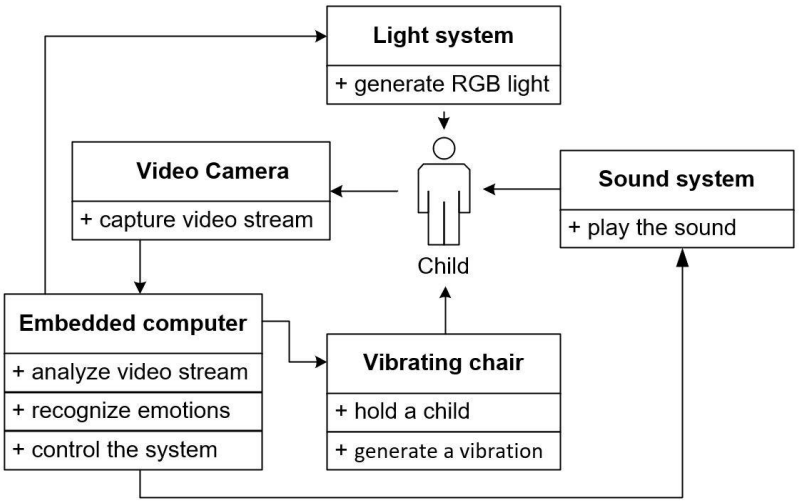


Figure 11. Schematic diagram of the Relaxation System
Source: created by the authors

The system's operation is governed by a control computer executing AI-driven software. This intelligent control software is engineered to orchestrate the vibrotactile chair functions, auditory-musical therapy, and chromotherapy, thereby generating individualized relaxation protocols. The software, leveraging artificial intelligence, executes a pre-programmed relaxation sequence and concurrently analyzes the subject's physiological and behavioural responses to the applied stimuli. Consequently, the AI system iteratively refines the relaxation protocol by learning the subject's preferential responses to specific light, auditory, and vibrotactile stimuli, as documented in (Kahou, S. E., et al. 2016). Furthermore, the AI software incorporates a critical event detection module, which alerts caregivers in instances of adverse reactions. An emergency cessation protocol is implemented to facilitate immediate termination of the relaxation sequence in unforeseen critical situations, such as pronounced negative responses to the therapeutic modalities (Fig. 12)

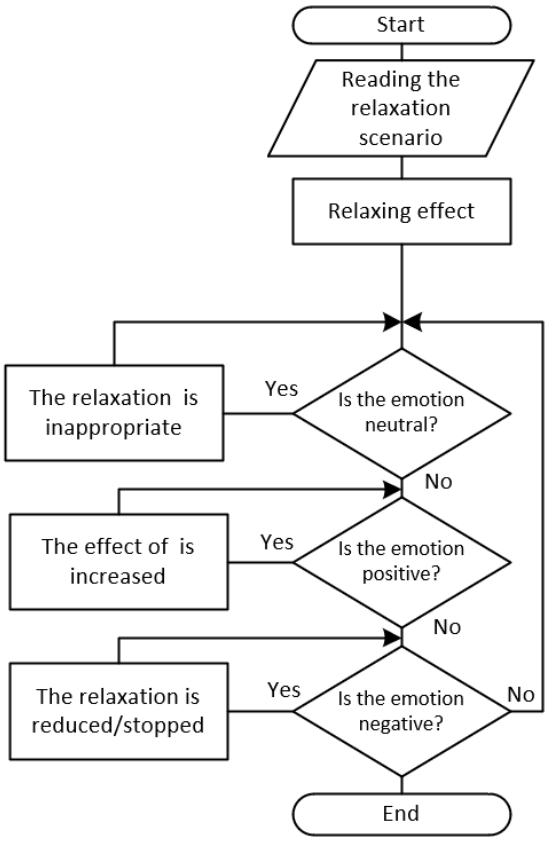


Figure 12. Algorithm for Controlling the Relaxation Scenario
Source: created by the authors

7. Hardware test of an Algorithm

An artificial intelligence algorithm, driven by recognized emotion models, operates by responding to changes in facial expressions (Zhang, T., et al., 2018), (Baltrušaitis, T., et al., 2012).

Initially, a predefined relaxation scenario created by a human is scanned. A relaxation effect tailored to the child is generated upon gathering the necessary data. If the child's face displays a neutral emotion, meaning the child's reaction to the stimulus is neutral, the parameters and intensity of the stimulus remain unchanged (Saini, S. S., & Rawat, P. (2022). The cycle continues until an emotional shift occurs. In the case of a positive emotion, the intensity of the stimulus is heightened while leaving other proportional parameters unchanged. Exiting the cycle is only possible when the child's emotional reaction to the relaxation scenario changes. Another logical condition awaits negative emotions, where the relaxation effect is reduced or, in specific cases, halted. Consequently, the system automatically responds to the child's agitation during relaxation. In the absence of emotional change, when negative emotions are absent, the cycle restarts from a neutral emotional state.

During each emotional cycle reassessment, relaxation scenarios can be adjusted by selecting the most optimal and acceptable ones for the specific child. This task is performed by another part of the program based on artificial intelligence algorithms – scenario selection, which is not elaborated upon in detail in this article.

Conclusions

The model didn't perform exceptionally well because it didn't have enough examples to learn from for certain emotions. It had a lot more data for "happiness" compared to "disgust," leading to a biased and less accurate model overall.

1. 64% Accuracy this is the overall performance of the emotion classification model.
2. Testing the model on arbitrary data not belonging to the FER dataset allowed for a qualitative assessment of the emotion recognition accuracy. It was revealed that due to the low resolution of the input image, an error occurs in the model's recognition.
3. Further research will be aimed at both improving the dataset used and developing the current convolutional neural network model.
4. The model is suitable for practical use in identifying positive, neutral, and negative emotions in children in relaxation system.

References

1. Alblushi, A. (2021). Face Recognition Based on Artificial Neural Network: A Review. Artificial Intelligence & Robotics Development Journal, 1(Issue 2), 116-131. https://doi.org/10.52098/airdj.202125 (seen 11.03.2024).
2. Baltrušaitis, T., et. al., (2012) 3D Constrained Local Model for Rigid and Non-Rigid Facial Tracking. IEEE Conference on Computer Vision and Pattern Recognition, CVPR.
3. Culjak, I., et al. (2012). A brief introduction to OpenCV. In 2012 proceedings of the 35th international convention MIPRO (pp. 1725-1730). IEEE.
4. Deva Priya W. (2022), Criminal Identification System to Improve Accuracy of Face Recognition using Innovative CNN in Comparison with HAAR Cascade. (2022). Journal of Pharmaceutical Negative Results, 218-223. https://doi.org/10.47750/pnr.2022.13.S04.023 (seen 11.03.2024).
5. Gaind, B. (2019), Emotion Detection and Analysis on Social Media, Global Journal of Engineering Science and Researches (IC'RTCET-18). 2019. - R 78-89.
6. Giannopoulos, P., Perikos, I., & Hatzilygeroudis, I. (2018). Deep learning approaches for facial emotion recognition: A case study on FER-2013. Advances in hybridization of intelligent methods: Models, systems and applications, 1-16.
7. Ioffe, S. (2020), Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift, 2015, https://aixiv.org/abs/1502.03167 (seen 11.03.2024).
8. Kahou, S. E., et al. (2016),''Emonets: Multimodal deep learning approaches for emotion recognition in video." Journal on Multimodal User Interfaces 10, No. 2: 99-111, 2016.
9. Sabiri, B., El Asri, B., & Rhanoui, M. (2022). Mechanism of Overfitting Avoidance Techniques for Training Deep Neural Networks. In ICEIS (1) (pp. 418-427).
10. Shukla, V., & Choudhary, S. (2022). Deep Learning in Neural Networks: An Overview. Deep Learning in Visual Computing and Signal Processing, 29-53.
11. Saini, S. S., & Rawat, P. (2022, April). Deep residual network for image recognition. In 2022 IEEE International Conference on Distributed Computing and Electrical Circuits and Electronics (ICDCECE) (pp. 1-4). IEEE.
12. Salman, S. (2020), Overfitting Mechanism and Avoidance in Deep Neural Networks, Salman. X. Liu arXiv.org. 2019. URL: https://aixiv.org/abs/1901.06566. (11.03.2024).
13. Viola, P. (2001), Rapid object detection using a boosted cascade of simple features, P. Viola, M. Jones, Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.-2001. - T. 1.
14. Zhang, T., et al., (2018), „Spatial-Temporal Recurrent Neural Network for Emotion Recognition ". IEEE transactions on cybernetics, (99), 1-9, 2018.

EMOCIJŲ ATPAŽINIMAS REALIUOJU LAIKU NAUDOJANT GILUMINIUS NEURONINIUS TINKLUS IR TAIKYMAS VAIKŲ RELAKSACINEI SISTEMAI

Eugenijus Mačerauskas[1], Maksims Žigunovs[2]
[1]Utenos kolegija,
Maironio g. 7, Utena, Lietuva
[2] Rygos technikos universitetas Liepojos akademija,
Liela iela 14, Liepaja, Latvija

Anotacija

Nagrinėjama emocijų atpažinimo iš veido atvaizdų, gautų iš vaizdo srauto, užduotis. Sprendimo metodas pagrįstas giluminių neuroninių tinklų naudojimu. Pateikiamas tinklo mokymui naudotas duomenų rinkinys, jo charakteristikos ir duomenų pasiskirstymas pagal emocijų klases. Aprašomi du konvoliucinių neuroninių tinklų modeliai: klasikinis šiai užduočiai sukurtas konvoliucinis neuroninis tinklas ir reguliarizavimo mechanizmais patobulintas konvoliucinis neuroninis tinklas. Remiantis gautais tinklo mokymo rezultatais, atliekama lyginamoji klasifikavimo tikslumo analizė. Aprašomas emocijų atpažinimo procesas naudojant bet kokius duomenis, nesusijusius su nagrinėjamu duomenų rinkiniu. Emocijų atpažinimo algoritmas buvo realizuotas aparatinėje sistemoje ir davo gerų rezultatų. Pateikiamas emocijų atpažinimo algoritmo praktinio taikymo pavyzdys relaksacinėje sistemoje skirtoje vaikams su emociniais suktrikimais.

Raktiniai žodžiai: emocijų atpažinimas, giluminis mašininis mokymasis, didelio tikslumo neuroniniai tinklai.