Thieme

Novel Method for Three-Dimensional Facial Expression Recognition Using Self-Normalizing Neural Networks and Mobile Devices

Neuartige Methode zur 3-dimensionalen Mimikerkennung durch den Einsatz von selbstnormalisierenden neuronalen Netzen und mobilen Geräten

Authors
Tim Johannes Hartmann[1,2], Julien Ben Joachim Hartmann[3], Ulrike Friebe-Hoffmann[2], Christiane Lato[2], Wolfgang Janni[2], Krisztian Lato[2]

Affiliations
1 Universitäts-Hautklinik Tübingen, Tübingen, Germany
2 Universitätsfrauenklinik Ulm, Ulm, Germany
3 Universität Stuttgart, Stuttgart, Germany

Key words
facial expression recognition, self-normalizing neural networks, facial geometry, disease recognition

Schlüsselwörter
Mimikerkennung, selbstnormalisierende neuronale Netze, Gesichtsgeometrie, Erkennung von Krankheiten

received 8.3.2022
accepted after revision 26.5.2022
published online 21.7.2022

Bibliography
Geburtsh Frauenheilk 2022; 82: 955–969
DOI 10.1055/a-1866-2943
ISSN 0016-5751

Georg Thieme Verlag KG, Rüdigerstraße 14,
70469 Stuttgart, Germany

Correspondence
Tim Johannes Hartmann
Universitäts-Hautklinik Tübingen
Liebermeisterstraße 25
72076 Tübingen, Germany
tim.hartmann@scanbooster.com

ABSTRACT

Introduction To date, most ways to perform facial expression recognition rely on two-dimensional images, advanced approaches with three-dimensional data exist. These however demand stationary apparatuses and thus lack portability and possibilities to scale deployment. As human emotions, intent and even diseases may condense in distinct facial expressions or changes therein, the need for a portable yet capable solution is signified. Due to the superior informative value of three-dimensional data on facial morphology and because certain syndromes find expression in specific facial dysmorphisms, a solution should allow portable acquisition of true three-dimensional facial scans in real time. In this study we present a novel solution for the three-dimensional acquisition of facial geometry data and the recognition of facial expressions from it. The new technology presented here only requires the use of a smartphone or tablet with an integrated TrueDepth camera and enables real-time acquisition of the geometry and its categorization into distinct facial expressions.

Material and Methods Our approach consisted of two parts: First, training data was acquired by asking a collective of 226 medical students to adopt defined facial expressions while their current facial morphology was captured by our specially developed app running on iPads, placed in front of the students. In total, the list of the facial expressions to be shown by the participants consisted of “disappointed”, “stressed”, “happy”, “sad” and “surprised”. Second, the data were used to train a self-normalizing neural network. A set of all factors describing the current facial expression at a time is referred to as “snapshot”.

Results In total, over half a million snapshots were recorded in the study. Ultimately, the network achieved an overall accuracy of 80.54% after 400 epochs of training. In test, an overall accuracy of 81.15% was determined. Recall values differed by the category of a snapshot and ranged from 74.79% for “stressed” to 87.61% for “happy”. Precision showed similar results, whereas “sad” achieved the lowest value at 77.48% and “surprised” the highest at 86.87%.

Conclusions With the present work it can be demonstrated that respectable results can be achieved even when using data sets with some challenges. Through various measures, already incorporated into an optimized version of our app, it is to be expected that the training results can be significantly improved and made more precise in the future. Currently a follow-up study with the new version of our app that encompasses the suggested alterations and adaptions, is being conducted. We aim to build a large and open database of facial scans not only for facial expression recognition but to perform disease recognition and to monitor diseases' treatment progresses.

ZUSAMMENFASSUNG

Einleitung Bisher beruhen die gebräuchlichsten Methoden zur Mimikerkennung auf 2-dimensionalen Bildern, obwohl es weiter entwickelte Methoden gibt, die 3-dimensionale Daten einsetzen. Diese benötigen aber stationäre Geräte, die weder tragbar sind noch im größeren Umfang bereitstehen. Da menschliche Emotionen, Absichten und sogar Krankheiten sich in spezifischen Gesichtsausdrücken oder durch Änderungen der Gesichtsmimik offenbaren können, ist eine kompetente und tragbare Lösung gefragt. Da 3-dimensionale Daten zur Gesichtsmorphologie eine höhere Aussagekraft haben und bestimmte Syndrome sich durch spezifische Gesichtsdysmorphien ausdrücken, kann dieses Problem dadurch gelöst werden, dass ein tragbares Gerät zur Erfassung von 3-dimensionalen Gesichtsscans in Echtzeit eingesetzt wird. In dieser Studie stellen wir eine neuartige Lösung für die 3-dimensionale Erfassung von gesichtsgeometrischen Daten und die darauf aufbauende Erkennung von Gesichtsausdrücken vor. Die neue Technologie, die hier vorgestellt wird, benötigt nur ein Smartphone oder ein Tablet mit integrierter TrueDepth-Kamera und erlaubt die Erfassung der Gesichtsgeometrie in Echtzeit sowie deren Zuordnung zu spezifischen Gesichtsausdrücken.

Material und Methoden Unser Ansatz bestand aus 2 Teilen. Zunächst wurden Trainingsdaten erstellt; dazu wurde ein Kollektiv bestehend aus 226 Medizinstudenten gebeten, bestimmte Gesichtsausdrücke anzunehmen, und ihre jeweilige Gesichtsmorphologie wurde währenddessen von unserer speziell entwickelten App auf iPads, die vor den Studenten aufgestellt waren, aufgezeichnet. Insgesamt bestand die Liste der Gesichtsausdrücke, die die Teilnehmer darstellen sollten, aus „enttäuscht", „gestresst", „glücklich", „traurig" und „überrascht". In einem zweiten Schritt wurden die neu erworbenen Daten dazu verwendet, ein selbstnormalisierendes neuronales Netz zu trainieren. Ein Satz aller Faktoren, die ein aktuellen Gesichtsausdruck zu einem bestimmten Zeitpunkt beschrieben, wird als „Snapshot" bezeichnet.

Ergebnisse Insgesamt wurden mehr als eine halbe Million Snapshots im Laufe der Studie aufgezeichnet. Im Endergebnis betrug die Gesamtgenauigkeit des neuronalen Netzes nach 400 Trainingsdurchgängen 80,54 %. Im Test betrug die Gesamtgenauigkeit 81,15 %. Die Sensitivität schwankte je nach Zuordnung des Snapshots und reichte von 74,79 % für „gestresst" bis 87,61 % für „glücklich". Bei dem positiven Vorhersagewert waren die Ergebnisse ähnlich, wobei „traurig" den niedrigsten Wert erreichte mit 77,48 % und „überrascht" den höchsten Wert erzielte mit 86,87 %.

Schlussfolgerungen Die Studie zeigt, dass respektable Ergebnisse erzielt werden können, selbst wenn anspruchsvolle Datensätze verwendet werden. Es wurden danach verschiedene Maßnahmen durchgeführt, die inzwischen schon in der optimierten Version unserer App integriert wurden. Damit sollten die Trainingsergebnisse voraussichtlich signifikant verbessert und in der Zukunft noch genauer werden. Zur Zeit wird eine Follow-up-Studie mit der neuesten Version unserer App durchgeführt, welche die vorgeschlagenen Änderungen und Anpassungen verwendet. Geplant ist nun der Aufbau einer großen, offenen Datenbank von Gesichtsscans, die nicht nur Mimik, sondern auch Krankheiten erkennen kann; damit könnten auch Fortschritte bei der Behandlung von Krankheiten verfolgt werden.

Introduction

Because facial expressions constitute a natural, powerful and universal form for humans to communicate intentions, physical sensations, and emotional states, their importance is signified [1, 2].

Significance of facial expressions in general and in diseases

Whereas the meaning of a certain facial expression is greatly influenced by the underlying situation [3], facial expressions deliver a vast amount of information about a person in general: Not only do a person's current emotions show in their facial expression [4] but also physical sensations felt by the person such as pain and even abstract concepts like intent [3, 5]. Even if an attempt is made to suppress certain emotional states or rather their expression, this usually does not succeed completely [6, 7, 8, 9]. Furthermore, can facial expressions serve as an important aid in diagnosing a multitude of different diseases: Mask-like facial expressions can be seen in Parkinson's disease [10], sad, expressionless, anxious faces in depression [11] and diminished facial expressions constitute a distinct domain of negative symptoms in schizophrenia [12, 13].

Currently, on the other hand, medical face masks limit the ability to express emotions through facial expressions as a substantial part of the face is hidden underneath them [14]. This can be especially limiting in the context of direct patient contact for both physician and patient: Because physicians receive limited information from their patients in regards to facial expressions, the assessment of the patient's current mental and emotional state may be limited [14]. Conversely, conveying emphatic expressions toward patients may be limited as well [14]. Summarized, medical face masks complicate social interaction as they disturb emotional

reading from facial expressions [14]. Of course, we do not want to discourage the use of medical face masks in today's challenging circumstances, but rather stress the need for both safe and reliable surrogates that allow people to express and read emotions.

Facial geometry

However not only facial expressions but also facial geometry can provide valuable information about patients and their diagnosis: For example, hypertelorism and further facial dysmorphisms are noticeable symptoms in a variety of syndromes, including LEOPARD syndrome [15], cri du chat syndrome [16] and Gorlin-Goltz syndrome [17].

Automated facial expression

Currently, as the field of artificial intelligence and machine learning is evolving rapidly, attempts are being made to classify images of persons and their facial expressions into distinct categories of the respective underlying emotions [18, 19, 20, 21, 22, 23, 24] or to objectively measure the severity of experienced pain [25, 26, 27, 28] or affect [29]. Until now, most research has relied on simple two-dimensional images as input data for training the artificial neural networks and consecutive evaluation [22]. Advanced approaches using two-dimensional videos promote advantages because videos additionally contain temporal information as well as further information on the shape of the participants' faces because the angle of view typically experiences subtle changes due to minimal movements of the head [22]. To train artificial neural networks and to evaluate their performance after the training, databases have been established [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. These databases contain two-dimensional images, image sequences, videos, or a combination of the former and show persons with different facial expressions [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44]. All of the data are categorized based on the respective expression [30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44].

To further enhance automated facial expression recognition accuracy through advanced data availability, databases have been created containing multiple images and/or videos from different viewpoints [37, 44] or even true three-dimensional data [45, 46, 47, 48, 49] of human faces and their respective facial expressions.

Facial expressions in current learning-environments

Since the emergence of the COVID-19 causing virus SARS-Cov-2 in 2019 [50], many new challenges have arisen all over the world [51, 52], including the field of medical research [53] and medical education [54]. Many educators have been confronted with new challenging requirements and the introduction of virtual classes [55]. Ultrasound programs have discontinued offering hands-on experiences to their students [56, 57], medical schools limited or suspended patient contact for students entirely [58] and some countries have even forbidden face-to-face university lectures completely [59, 60]. With the ongoing pandemic, new tools such as apps and services for virtual online meetings have surged and become a new standard in current medical teaching environments [51, 61, 62, 63, 64, 65] and diagnostic relations [66]. These tools and solutions posed a viable surrogate for the first time in the pandemic because they allowed a continuation of the teaching process and have proven to be at least equivalent to traditional approaches in earlier studies [51, 59, 67].

Furthermore, it has been shown that the introduction of Zoom virtual tutorials resulted in higher student satisfaction and a reduction in instructor workload of approximately 25 %, whereas engagement levels of students and the grade distribution stayed the same [68]. To take these new chances and opportunities, it becomes clear that automated recognition of the learner's status and the corresponding adaptation of teaching and learning requirements must be taken into consideration. Since facial expressions can reflect both excessive and insufficient demands on learners as well as their comprehension, automated facial expression recognition can also play a major role here, detecting the respective state of each student and consecutive adaption of the learning material through the teaching software [69].

Taking into account the multitude of applications that automated facial expression recognition allows, and the possibilities concerning the detection of pathologies through automated facial scans, the need for a technology that permits not only facial expression recognition but also three-dimensional facial scanning becomes evident. As real-time three-dimensional facial scans can be utilized both as a means to achieve automated facial expression recognition and aid in detecting pathologies, a highly portable, both flexible, and cost-efficient technology to swiftly perform three-dimensional facial scans would be desirable.

In this study, we want to present a highly portable and flexible solution to perform automated facial expression recognition and acquisition of three-dimensional facial geometry data using only pre-existing smartphones or tablets to cut costs. Using our setup, the only requirement of the devices is to be equipped with a TrueDepth camera.

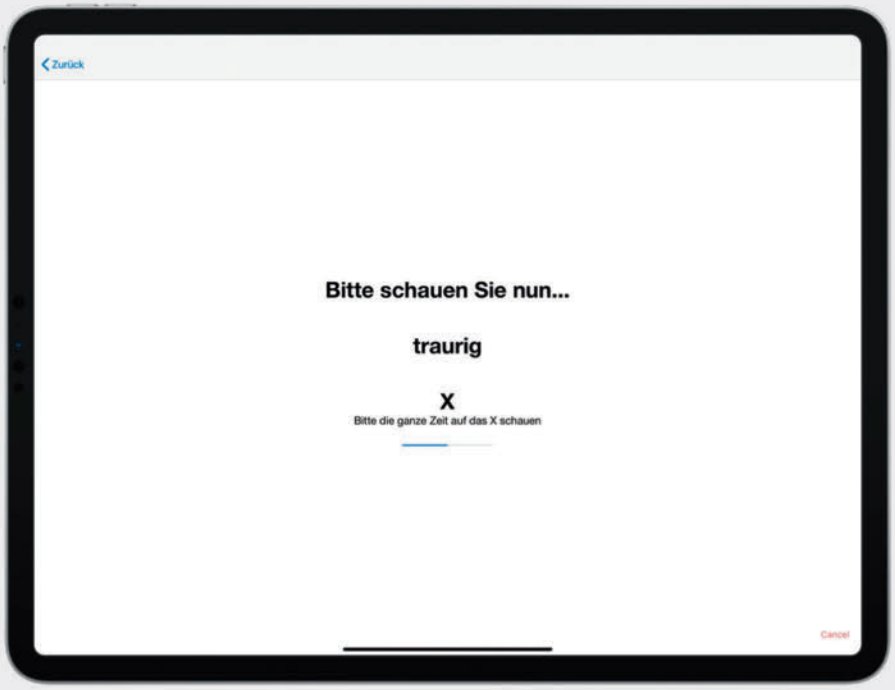
In the study presented in this work, this technology was implemented – to our knowledge – **for the first time** using only widely available technology.

Since the devices used could even be the patients' own, large databases could be created quickly, which would enable seamless data acquisition during treatment periods, for example.

Material and Methods

A collective of 226 medical students from Ulm University in the 8th–9 th study semesters that were present during the block internship in gynecology and obstetrics between April and December 2019 were examined. The study examination was integrated into the ultrasound seminar with an average of 10 participants per appointment. All students who regularly attended the seminar or block internship were included, provided they had given their consent to (voluntary) participation.

Before the study was carried out, the respective participants were informed about the course of the study, its voluntary nature, and the absence of any (negative) consequences in the event of non-participation. According to a request to the Ethics Commission in Ulm, this study fell under § 15 of the professional code for physicians in Baden-Württemberg and therefore did not require an ethics vote.



▶ **Fig. 1** The screen that the participants were shown during the examination: It was equipped with instructions on the facial expression to obtain ("Bitte schauen Sie nun traurig" – "Please look sad now"), an "X" which indicated the preferred position on the screen to look at as well as the instructions therefore ("Bitte die ganze Zeit auf das X schauen" – "Please look at the X all the time").

Setup

At the beginning of the study, the participants were seated at tables with one third-generation Apple iPad PRO per student placed on the table's surface. The iPads had a screen size of 12.9″ and were running our newly developed app to perform the facial scans. Each of the iPads was equipped with a SmartFolio cover, which was used to set it up on the table. The angle between the table's surface and the back of the iPad was approximately 60° so that the device's front and the TrueDepth camera of the device pointed directly at the participant's face.

Description

In total, the list of the facial expressions to be shown by the participants consisted of "disappointed", "stressed", "happy", "sad", and "surprised". On the iPads, the screen shown in ▶ **Fig. 1** was displayed, containing the current expression to be shown by the participants in its center.

In the background, the TrueDepth camera of the iPad was now activated, and the recording process started.

The TrueDepth camera is a hybrid camera system that consists of both an ordinary camera and a projector, which projects around 30.000 points of light onto the subject's face and subsequently records their reflection and the time-of-flight difference using an infrared camera [70]. The functionality is thus similar to a lidar (light detection and ranging) [70]. This means that an exact three-dimensional scan of the human face can be recorded in real-time by the system [70]. The raw data now recorded was automatically analyzed using the ARKit framework and an ARFaceAnchor object was subsequently created [71, 72]. Utilizing the newly created ARFaceAnchor object, extraction of the current facial geometry and current facial expression from it was possible. The facial geometry was provided in form of an ARFaceGeometry object as "a coarse triangle mesh representing the topology of the detected face" [73]. The facial geometry however was not processed further in this study. The facial expression was represented in form of a dictionary containing the type "blendShapeLocation" matched with a floating-point value [74, 75]. In general, a dictionary is a programmatic type that contains an array consisting of a certain key and a value matched with it (key-value pairs) [76]. In our example, the dictionary contained an array of parameters relating to components of the facial expression and their relative strength. There were 52 of these parameters ("blendShapeLocation") contained for each facial expression recognized along with their respective strength. The relative strength was described by floating-point values that could adopt values between 0 and 1 [75]. The values for the relative strength were subsequently saved, accompanied by the type of facial expression to be shown as well as the current time.

In the following, the totality of all 52 parameters recorded at a specific moment shall be referred to as a "snapshot".

It was checked by our software once per second if the minimum number of 500 such individual snapshots (each with 52 specific features of the face) had yet been reached for the current type of facial expression. If the minimum number had been

reached, the next type of facial expression was displayed on the iPad's screen and the participants were asked to adopt it. The current progress was indicated by a bar in the middle of the iPad's screen. During the whole examination, a point marked with an "X" was displayed in the screen's central area. The participants were asked to preferably look at this point. This was integrated to prevent the students from looking away from the iPad, which could lead to a deterioration of the recognition quality or even complete transient loss of the system's recognition. If the detection through the camera failed (e.g. triggered by the student looking away from the camera or temporarily poor lighting conditions), the acquisition was automatically stopped. This prevented invalid values from being recorded.

Programming

The snapshots obtained were consecutively saved in a file called "EmotionTrain.txt", combined with the respective facial expression to be displayed and the time of recording.

AI

The data obtained were used to train a new type of ANN architecture.

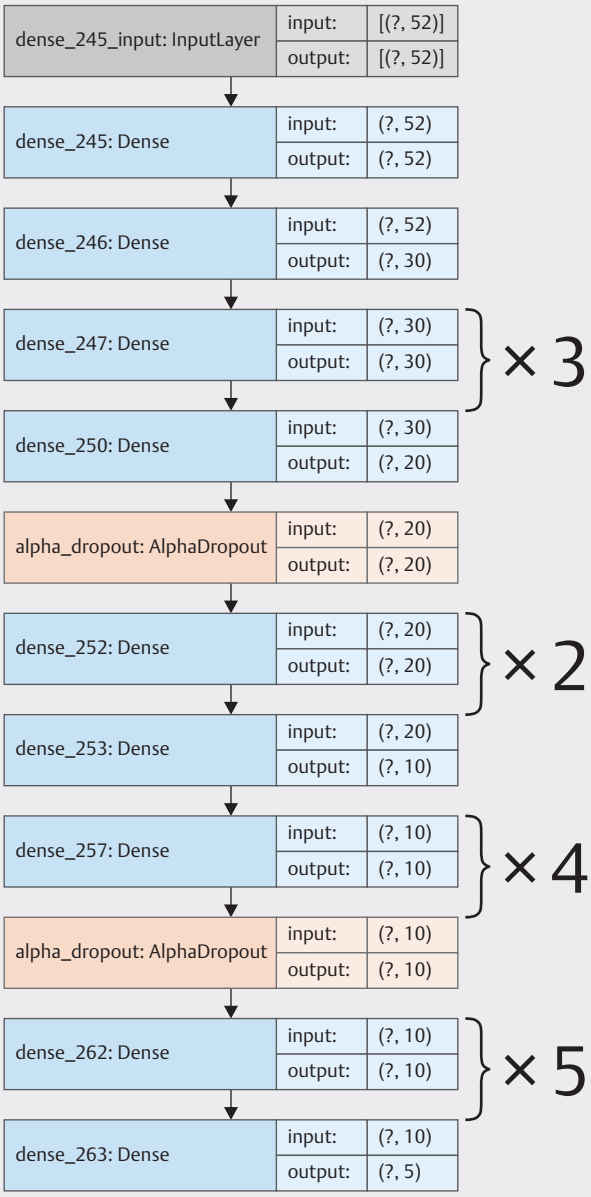
The study was designed in such a way that a minimum number of around 500 such snapshots per type of facial expression had to be recorded to continue to the next type of expression because the resulting data set was supposed to be balanced in terms of the number of snapshots per facial expression. As this was a minimum number, which was checked by a program function at regular intervals and, when exceeded, the next facial expression or finally to the end of the study, the total number of registered snapshots during the training part is not an integral multiple of the product of study participants and the number of categories of facial expressions. The normal distribution of the snapshots per category was checked statistically.

The recorded snapshots were separated into two sub-sets: A training-set, and a test-set. The training-set was used to train the network and consisted of 85% of the data set's total amount of snapshots, whereas the test-set consisted of the remaining 15%. The test-set was presented to the network only after the training was fully completed. It was used to assess the network's accuracy and specifics.

Allocation of the snapshots to the respective set took place utilizing a random generator and after shuffling the entire data set [77].

An artificial neural network (ANN) was then created using the environment TensorFlow [78]. The ANN contained a total of 22 different layers, including an input layer, an output layer, and the 20 hidden layers in between (see ▶ **Fig. 2**). These are made up of dense layers and dropout layers. The dropout layers were used to minimize the overfitting of the network [79].

In addition, an L2 regularization of 0.25 was used for some of the dense layers to further minimize overfitting [81]. This limits the weights, determined for the corresponding layer depending on the sum of the square of the weight. Hereby, extreme value formations of the factors are minimized [81]. Since the input values of the ANN were floating-point values in the ranges from 0 to 1 and these should be processed accordingly in the network, scaled exponential linear units (SELU) were used as the activation function of the hidden layers [82]. The resulting ANN has subsequently been trained using Adam optimization. This was chosen because it requires little memory, can be calculated efficiently, is well suited for problems that have a large amount of data and parameters, and remains unchanged when the gradient is rescaled diagonally [83].



▶ **Fig. 2** Schematic representation of the network architecture, based on the function plot_model of the environment TensorFlow/Keras [78, 80]. The input layer is shown in the first box. This is followed by various levels of the "Dense" and "AlphaDropout" layers. Finally, the output layer is shown. If a curly bracket was shown next to a box, this means that the corresponding section is repeated by the number given.

Special network architecture Primarily, an ANN of the Self Normalizing Network (SNN) subtype was used. This type of ANNs is characterized by the fact that no a-priori normalization, i.e. conversion of the data, which have different scales, to a common, defined scale from 0–1, needs to take place before training [82, 84]. For an ANN to achieve self-normalizing behavior, a SELU was used as an activation function, which was first published in 2017 by Klambauer et al. [82]. Since the data used were already normalized, no further normalization would have been necessary. Concludingly, a conventional, non-self-normalizing network could have been used. The raison d'être of the related SNN is that SNNs have further advantages compared to conventional, non-self-normalizing network architectures: They allow shortened training times, faster convergence, and a high degree of robustness in training [82].

When using SNNs, however, it is mandatory to use special initialization functions [85]. In our case, we used the LeCun normal function, which was published by LeCun et al. [86].

Furthermore, when using SNN it is mandatory to use a special subform of the dropout layers: Alpha dropout layers [82, 85]. With these, the activation is not simply deactivated (set to zero), but provided with the negative saturation value of the SELU activation function [82, 87]. This means that the self-normalization property of the SNN is retained even after the dropout. Klambauer et al. were able to show in 2017 that this method is superior to the usual use of regular dropout layers when using an SNN [82].

A relatively low dropout rate of 0.1 was chosen since tests showed that such lower rates still reliably limit overfitting, but concurrently allow a high level of network accuracy.

Inference test

After completion of the training, a test to determine inference time was conducted. Inference denotes the application of an artificial neural network. In inference, no training takes place. Instead, data is supplied to the neural network to receive the wanted predictions.

For this, a standalone app was written in Swift, runnable on both iOS and macOS. The fully trained network was converted into a TFLite network and implemented, as were the test- and training-sets. To measure the inference time, the current system time was captured both before and after calling the invocation of the interpreter (interpreter.invoke()). To capture the system time, the command CACurrentMediaTime() was executed. Inference time was measured for the data of the test-set, training-set, and a total of 500 000 randomly generated floating-point precision (Float32) values between 0 and 1. These were generated by the command Float32.random(in: 0 ... 1). The inference test app was executed on a 2021 MacBook PRO with an M1 Max processor, an iPhone 12 PRO with an Apple A14 Bionic processor, and an iPad PRO 12.9″ 3 rd generation with an A12 X processor. During the test, only the test app was actively running on the devices. They were connected to a permanent power connection, and the battery charge levels were at least 50 % in each case.

Statistics

The descriptive statistics were carried out by specifying absolute and relative frequencies for categorical data. Specifics to describe the performance of the artificial neural net, its training and test included recall, precision, F1-Score, and specificity for each category. The equal distribution of snapshots per category was checked using a χ^2 test.

The statistics program IBM SPSS Statistics for Windows, Version 25 (Armonk, NY: IBM Corp.), as well as Microsoft Excel was used for all statistical analyzes.

Results

Results from the data acquisition

In total, over half a million snapshots were recorded in the study (n total data set = 563 226).

These snapshots were separated into two sub-sets: a training-set and a test-set. The training-set was used to train the network and consisted of 85 % of the data set's total amount of snapshots (n training-set = 478 742), whereas the test-set consisted of the remaining 15 % (n test-set = 84 484).

The training-set consisted of snapshots with five different categories, which are shown in ▶ **Table 1**.

It can be seen here that the individual categories are each composed of an average of 95 748.2 individual snapshots. The χ^2 test carried out confirmed an equal distribution. The quantitative differences between the individual categories per set have their origin in the random programmatic selection when creating the training and test-sets. The homogeneous composition of a training-set is of fundamental importance for the type of ANN presented here.

As already described above, the test-set consisted of 15 % of the total data set.

In ▶ **Table 2** the distribution of the number of snapshots per category is shown. They are equally distributed as well.

▶**Table 1** Composition of the training-set from the various categories, as well as χ^2 test to check the equal distribution.

–	Disappointed	Stressed	Happy	Sad	Surprised	Total number
Number of snapshots in training	95 402	95 585	95 734	96 270	95 750	478 741
Frequency	19.93 %	19.97 %	20.00 %	20.11 %	20.00 %	–
χ^2 test	0.357 537 338					

▶**Table 2** Composition of the test set from the various categories, as well as χ^2 test to check the equal distribution.

–	Disappointed	Stressed	Happy	Sad	Surprised	Total number
Number of snapshots in training	17086	16876	16781	16972	16768	84483
Frequency	20.22%	19.98%	19.86%	20.09%	19.85%	–
χ^2 test	0.372682717					

▶**Table 3** Listing of the specifics of the individual categories during a run with all data of the training set.

By category	Disappointed	Stressed	Happy	Sad	Surprised
Recall/Sensitivity	80.917%	76.911%	88.245%	83.006%	81.381%
Precision/Positive predictive value	81.774%	79.520%	83.582%	78.487%	87.549%
F1-Score	0.81342859	0.78193721	0.85850169	0.8068335	0.84352166
Specificity	95.512%	95.059%	95.667%	94.273%	97.106%
Negative cases	383339	383156	383007	382471	382991
Total					
Recall/Sensitivity	82.095%				

▶**Table 4** Listing of the specifics of the individual categories during a run with all data of the test set.

By category	Disappointed	Stressed	Happy	Sad	Surprised
Recall/Sensitivity	80.118%	74.793%	87.605%	82.842%	80.439%
Precision/Positive predictive value	80.986%	78.252%	82.697%	77.483%	86.868%
F1-Score	0.805495896	0.764830637	0.850801551	0.800728971	0.835299582
Specificity	95.231%	94.811%	95.457%	93.948%	96.989%
Negative cases	67397	67607	67702	67511	67715
Total					
Recall/Sensitivity	81.152%				

The test-set was used to assess the network's accuracy and specifics.

Allocation of the snapshots to the respective set took place employing a random generator and after shuffling the entire data set [77].

Results from the training of the ANN

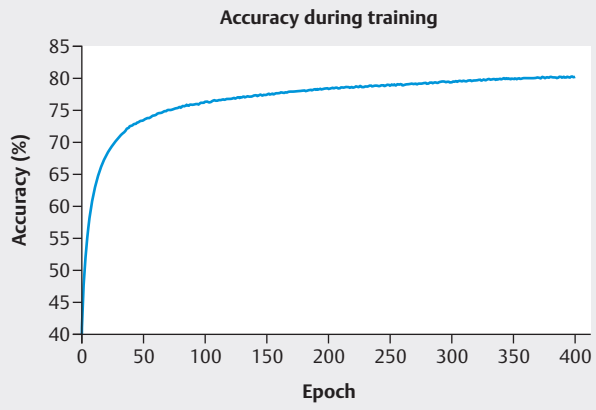
The training of the network was accomplished within a day using a high-performance system (Google Tensor Performance Unit 2). The overall achieved accuracy of the ANN's predictions was 43.58% at the start of the training. It rose to 80.54% over the course of the training. The exact course of the accuracy and the loss function during the training can be seen in ▶**Fig. 3** and ▶**Fig. 4**. After 400 epochs, the training was stopped. The batch size used was 128.

Results after the training

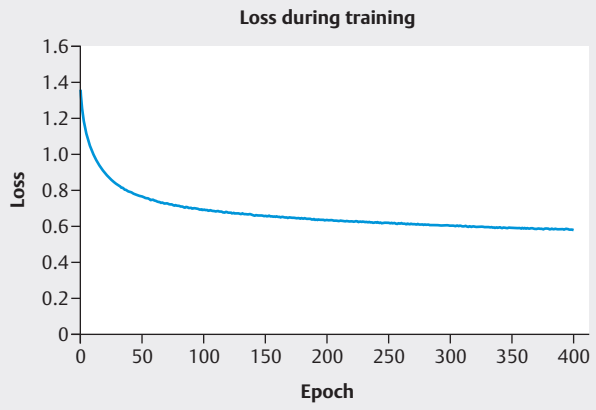
Results with the training-set

When the training-set was used once again in a subsequent evaluation run, an overall accuracy of 82.095% could be determined for all snapshots of this set. The ANN was not trained in this evaluation run – the training-set was only used as the input data set. Precision, recall, the F1 score, and the specificity of the individual categories, as well as the overall accuracy across all categories, are shown in ▶**Table 3**.

In this evaluation run, overall accuracy was achieved that exceeded the one determined during the training although the utilized data were identical.



▶ **Fig. 3** Representation of the accuracy achieved during training depending on the number of passes (epochs).



▶ **Fig. 4** Representation of the values of the loss function achieved during training depending on the number of passes (epochs).

Results with the test-set

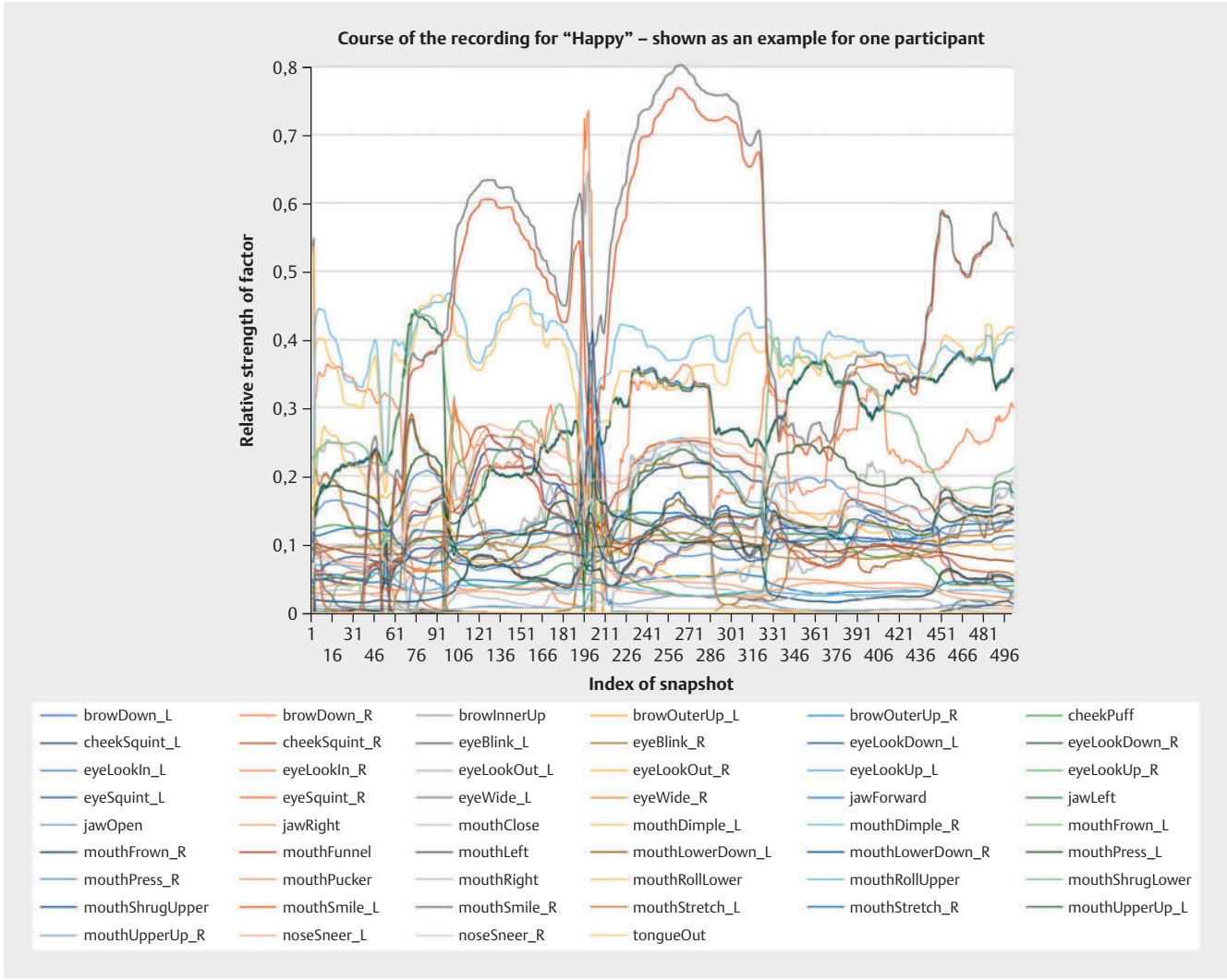
Using the test-set, an overall accuracy of 81.152 % was determined for all snapshots of this set. The performance data and sensitivities of the individual categories are shown in ▶ **Table 4**.

Inference results

Inference times resulting from the application of the network for a single inference cycle are given below. They ranged between 9.71 microseconds (µs) on the fastest tested device and 1338.67 µs on the slowest device. Each longest inference duration per category and device differs from the average duration by a factor of 10 or higher. During the test, inference of the artificial neural network utilized one processor core out of six on the iPhone 12 PRO and out of ten on the MacBook PRO. The resulting reachable mean facial expression recognition rates ranged between 42 283.29 per second on the iPhone 12 PRO and 90 090.09 on the MacBook PRO on a single core.

▶**Table 5** Inference times needed for a single prediction cycle using the final trained artificial neural network on different devices. Input data was supplied in form of the test- and training-set as well as random numbers. The listed numbers denote the inference times in microseconds (µs).

	Number N of supplied snapshots	Range of values in µs	Minimum inference time in µs	Maximum inference time in µs	Mean inference time in µs	Std. Deviation in µs
MacBook PRO M1 Max random	500 000	128.46	9.71	138.17	11.16	0.75
MacBook PRO M1 Max test-set	84 483	106.42	10.08	116.50	11.11	0.71
MacBook PRO M1 Max training-set	478 741	114.12	10.00	124.12	11.10	0.76
iPhone 12 PRO test-set	84 483	618.62	11.50	630.12	20.20	12.25
iPhone 12 PRO training-set	478 741	408.12	11.54	419.67	22.93	14.43
iPhone 12 PRO random	500 000	744.54	13.67	758.21	23.65	15.36
iPad PRO 12.9″ 3 rd gen test-set	84 483	1323.63	15.04	1338.67	18.41	5.36
iPad PRO 12.9″ 3 rd gen training-set	478 741	1225.54	14.63	1240.17	18.42	3.34
iPad PRO 12.9″ 3 rd gen random	500 000	762.33	14.83	777.17	18.92	3.19



▶ **Fig. 5** Example representation of the course of the raw data of a participant for the category “Happy” during the recording of the data for AI training. The names of the individual factors listed in the legend are in accordance with [74].

Discussion

Training and evaluation of the ANN

Looking at the previously presented results, both problematic and positive aspects can be found. The various aspects are to be classified below.

Discussion of negative aspects

1. Inhomogeneity:
   - Since the study participants were asked to show the facial expression displayed on the screen, but checking whether the study participants followed this command over the entire period (i.e. until the minimum threshold of 500 snapshots per facial expression was reached) was impossible, the data may show a certain inhomogeneity.
   - ▶ **Fig. 5** shows an example of the recording of the facial expression “happy” of a participant over the entire period. It can be seen that the factors “mouthSmile_R” and “mouthSmile_L” reach the highest values. Such a result is to be expected since these two factors represent the relative strength of the mouth’s left and right corners’ elevation, which is typical for smiling. However, it can also be seen that the relative strength is subject to certain fluctuations and even temporarily low values near zero.
   - Especially at the start of the measurement, it should be noted that the rise of these values requires a certain amount of time, namely that which is necessary for the participant to react to the new command on the tablet display. Since the recording of the snapshots took place at an approximate rate of 30 s^{-1}, the time needed to react to the modified command took about 1–2 seconds for this participant.
   - The values of these curves also decrease towards the end, which can be explained, for example, by a lack of or declining motivation of the participants. It must also be noted that facial expressions with emotional connotations can be represented in a variety of ways, of course, both intra- and interculturally as well as individually differing [2].

- These specifications make it difficult for an ANN to develop generally applicable algorithms, especially for previously unknown participants. In addition, they reduce the predictive quality of the ANN for the data of the known participants, since the facial expressions may not be recorded completely homogenously because of the one to two-second overlap time between the individual categories.

2. Disparity between emotions and facial expression:
   - In addition, it must be noted that the emotions that a participant felt during the study do not necessarily result in a change of the facial expression or a facial expression per se [4].
   - Vice versa, it must be noted that the participants were asked to adopt certain facial expressions without necessarily being in the same emotional state that would cause the same expression independently from laboratory conditions. Therefore, it must be questioned if the facial expressions collected under a certain category are really of this category and would look the same if a participant truly felt happy, sad, and so on or if the shown facial expression only represented the idea of the participant how a facial expression of the respective category would show.
   - This leads to a reduced informative value of the ANN's results.
   - To help solve this issue in future studies, media could be shown which have an increased probability of triggering corresponding facial expressions due to their emotional connotation. For validation purposes, each participant could subsequently be asked which emotion (or whether at all) he or she felt during the playback of, for example, a short movie clip. Afterward, the ANN should be retrained using the resulting data.
   - Furthermore, this ANN should demonstrate a fundamental principle using the latest and at the same time widely available technologies (AI and smartphones/tablets with TrueDepth camera).
3. Course over time:
   - The snapshots were recorded over time but were used and evaluated individually. It can therefore be difficult to infer the facial expression shown from snapshots, as this may only be revealed by looking at their course over time.
4. Baseline:
   - No "baseline" facial expression (neutral facial expression) was recorded during the study because during planning it was assumed that such neutral expressions would be reflected in homogeneously low values for the individual factors.
   - After careful consideration and under the assumption of the above hypothesis that facial expressions can show a significant intra- and intercultural as well as individual variability, it can be determined that the inclusion of a baseline category for future questions should be contemplated to offer even better possibilities in assessing differences between participants.
5. Snapshots
   - The used factors, automatically generated by ARKit and in their entirety forming a snapshot, remain undefined both in terms of their generation, reproducibility, and significance.
     - The documentation of the ARKit framework used merely notes that these reflect the particular specifics of the facial expression currently shown.
     - How this dimensional reduction from the raw data, which is based on the values registered by the TrueDepth camera, to just about 50 floating-point values takes place, remains concealed.
     - How precise the recognition of the specifics of the facial expression is, based on the raw data and how exactly – in reverse – a facial expression can be defined at all based on these 50 floating-point values, is not explained either.
     - Whether the same or similar facial expressions produce the same or similar snapshots also remains undefined.
   - However, two factors must be considered: First, according to documentation, these factors were designed to animate virtual characters based on faces recognized by the camera. This leads to the conclusion that the reproducibility must be given, otherwise, an animated face would show e.g., clear twitching while a constant facial expression is being shown. Secondly, the high quality of the recognition and the sufficiently large information content of the approximately 50 factors can be seen in the fact that overall accuracy of the network of approximately 80 % could be achieved (in the test). If there had been poor recognition of the factors, low reproducibility, or insufficient information content of the factors, no accuracy to this extent would have been achievable as the supplied data would have been at least partly random.
6. Dimensionality reduction:
   - The goal of the AI used here is inevitably the dimensional reduction of the acquired data: from many thousands of measured values of the TrueDepth camera or subsequently the snapshots with almost infinite value possibilities, statements with only one dimension and few finite possibilities are to be distilled.
   - By reducing information from many thousands of light points projected onto the face by the TrueDepth system to just about 50 factors, the network is relieved of a step it would have had to take inevitably (dimension reduction), but using raw data for training might have yielded even more accurate results since the 50 factors ultimately follow at least partly arbitrary goals:
     - It is not proven that the factors used already describe each facial expression sufficiently precisely that the category of the same can be extracted from them, or formulated differently, that 100 factors would have been possible as well – or even necessary.
     - The network could have been trained differently, yielding different target factors.

- It is therefore questionable whether these about 50 factors truly represent the 50 most optimal, efficient factors for our purpose.
- Concludingly, approaches should be evaluated that take the entirety of facial geometry as an input.

7. Selection of training data:
   - For the training, all snapshots of all participants were considered together and divided into two sets (training-set and test-set).
   - Since the factors were recorded with a sufficiently high frequency and it is to be expected that the measurement conditions (lighting conditions, position of the participant's head, and facial expression) sometimes did not change greatly between two snapshots. Therefore, some of the snapshots used in the training may have similarities to those used in the test.
   - Thus, it is to be feared that the results reflect falsely high values that could not have been achieved by using completely independent data, as would have been the case, for example, if the data had been assigned to the respective sets, based on participants.
   - If, however, the allocation of data into training and test-sets had not been based on individual snapshots, but on participants (a participant or all of his snapshots are allocated to either one or the other set), the network would have had the same amount of data available for training, but a significantly greater heterogeneity between the data of the test and training-sets would have to be expected. This is especially critical given the small number of participants (226) since in essence only 192 (226 × 85 %) different examples of human behavior would have been available to the network for training.
   - Since the entire spectrum of human facial expressions can hardly be shown based on only 192 persons during a half-hour investigation, this procedure would certainly not have resulted in a robust network capable of making general statements.
   - On the contrary, it would have to be feared that some of the participants assigned to the test-set could not have been categorized sufficiently correctly based on the snapshots of the 192 persons available for training.
   - It must also be considered that there are intercultural differences between the perception of a facial expression and the composition of facial expression as a reaction to a certain situation. Similarly, interindividual differences must be taken into account when questioning whether all participants could have been analyzed accurately with a participant-based assignment of snapshots.
   - Considering the goal of the network in this pilot study to at least categorize the facial expressions of the participants present in this study with sufficient accuracy, a single-snapshot-based mapping seems to be better, since thus the previous limitations can be circumvented.
8. Conclusion:
   - Finally, several conclusions can be drawn from the previous observations described here, which should be used in future examinations and studies:
     - There should be a clear separation of facial expressions, in which the participants should be specifically advised to show their facial expressions for the entire time.
     - In addition, the participants should be able to start the recording with a start button as soon as they have adopted the new facial expression.
     - It should be considered not only to use individual snapshots as input for a revised ANN but rather a series of temporally contiguous snapshots, since it may not be possible to fully deduce which facial expression a participant is showing from a single snapshot.
     - In the future a new category "neutral facial expression" should be added to the training.
     - To be able to make more general statements, a significantly larger and more diverse population of participants would have to be used.

Advantages

The analysis shows that the training of the ANN, despite the use of data with difficulties (see above), delivered results. After all, an overall accuracy of 80.54 % was observed at the end of the training and even higher accuracy when using the test-set, which was previously completely unknown to the ANN.

Of course, the question arises as to how it can be explained that the accuracy when using the test-set exceeded that of the training-set. To do this, the three-stage process must first be reiterated: Initially, the network was trained, at the end of which the overall accuracy in the prediction was 80.45 %. This was followed by an evaluation run in which a higher accuracy of 82.095 % was achieved using the same data that was used for the training (the training-set) again. Finally, there was an evaluation run with data that was completely new for the network: the data from the test-set. Here again, a slightly lower accuracy of 81.152 % was achieved.

Because the same data was used in steps one and two, it would be expectable that the same results in terms of accuracy would ensue. There are several factors however that explain why different results arise from the same data: First of all, several alpha dropout Layers were used, which lead to the deactivation of randomly selected neurons during training. This is done by setting the weight of the respective neuron to zero when using dropout layers and by assigning a negative value for alpha dropout layers, a subtype of the dropout layer [79, 87]. This of course limits the capacity and complexity of the ANN. After the training – as soon as the ANN is used to make predictions (as in the subsequent evaluation run) – this behavior (zeroing or negative value assignment) is switched off, whereby the ANN, in abstract terms, receives previously unavailable calculatory capacities.

In summary, the alpha dropout layers are activated during the training and thus lead to a reduction in the complexity and thus the calculation ability of the ANN. They are deactivated in the evaluation process, which means that the ANN has more calculatory units available, and the complexity of the ANN increases. Due to the increased complexity of the ANN, it was able to make more precise statements in our case.

If one now considers that the overall accuracy of the training-set in the evaluation cycle is superior to that achieved during training with the same data, the influence of the dropout layers can be determined. The minimally worse result when using the data from the test-set in the third run compared to the result in the second run using the data from the training-set is around a 1 % difference in absolute terms. The poorer performance when using new data unknown to the network (in our case the data from the test-set) is called overfitting. With only around 1 % absolute difference, however, only minimal overfitting of the ANN has taken place. Therefore, our strategy of using dropout was beneficial.

Another factor that contributed to the observed behavior is the L2 regularization we used. This limits the choice of weights since extreme value formations are penalized [79, 88]. Since this is intended to minimize overfitting, it is also minimized that the test accuracy decreases compared to the training accuracy.

To ultimately determine that the observed behavior did not result from an "advantageous" distribution of the snapshots of the test-set, k-fold cross-validation should be carried out in future investigations [89]. In the current study, however, this was not implemented because the assignment of snapshots to either set had taken place randomized. Additionally, if the standard value of k = 10 had been used for the k-fold cross-validation, a total computing time that would have surpassed our available capacities would have ensued on the utilized system.

Selectivity of the categories

Finally, when looking at the individual categories set out in ▶ **Table 4**, it can be seen that the values achieved for the sensitivity differed significantly in some cases. The difference between "stressed" and "happy" is almost 13 % when using the test-set. In the training-set, too, these two categories were those with the highest and lowest values for sensitivity and therefore the furthest apart. Ultimately, the different values can be explained, by the fact that the category "stressed" has a significantly lower degree of distinction or significantly fewer distinguishing features in relation to the other categories, whereas the category "happy" does. This would mean that the ANN would have a lower probability of correctly classifying a "stressed" snapshot as such. This also explains the lower positive predictive value compared to the other categories. If a snapshot is assigned as "stressed", the probability is now lower, since other categories sometimes have similar characteristics (there is a low degree of selectivity).

Real-world application – inference

As shown in ▶ **Table 5**, the inference times ranged between 9.71 µs on the fastest tested device and 1338.67 µs on the slowest device. Sometimes, great differences in duration time occurred between the inference cycles, resulting in a difference from the average duration by a factor of 10 or higher. These can be explained by a power-saving state of the processor. When inspecting the data, it was noticeable that the first processed snapshot in particular always took much longer than the average. Before the processor-intensive process of inference begins, it can be assumed that the processor load is significantly lower. It is therefore likely that the processor was throttled at this point to save energy (e.g. by reducing the clock frequency). Furthermore, some components such as variables are initialized in the first pass, which costs additional time and is omitted in subsequent passes.

Considered on its own, average processing rates of several tens of thousands of snapshots per second were possible in the conducted tests – a multiple of what is needed. However, it must be considered that in real operation, other essential processes are operating concurrently: For example, the three-dimensional image must be calculated and generated. Likewise, the 52 factors must be determined. Accordingly, fewer resources are available in real operation for inference than were in the test carried out to determine the inference times. Nevertheless, the low, measured inference times show that only about 0.75 ms of processor time are demanded for processing 30 snapshots, based on a mean inference cycle duration of 25 µs. This corresponds to a theoretical mean temporal processor occupancy of 0.075 % which provides a reasonably large buffer to allow a problem-free execution in real-time under real conditions. Consequently, this new technology allows even old devices to perform a real-time evaluation of 3 d facial expression recognition with ease. Additionally, it can be assumed that the inference rates under real conditions are even higher since the test app was executed and controlled via the Xcode development environment, and debug information was collected in the process.

Conclusion

With the present work, it can be demonstrated that respectable results can be achieved even when using data sets with some challenges. The ANN used turned out to be very robust in the evaluation, with special reference to the above-described challenges.

Furthermore, the use of the latest and at the same time widely available technologies is to be classified as very interesting for future projects. All that is required for the setup presented is an iPhone or iPad with a built-in TrueDepth camera. These system requirements can already be found in devices that are now more than 4 years old (e.g. iPhone X, presented in September 2017) [90].

Outlook

Use in other areas would therefore be classed as quite cost-effective, simple, and flexible. A binary, ordinal, or even cardinally scaled classification for automated pain recognition of (hospitalized) patients based on their facial expressions, automatic adaptation of the learning content on tablets, depending on the facial expressions shown by the students, or observation of the effects of interventions on patients during treatment periods would be possible. Depending on the reaction to the content shown on the tablet or smartphone, an automatic adjustment could take place in real-time.

Implications for further examinations

As described, significantly more and more complex data should primarily be recorded in a subsequent study: Not only should the 52 different factors relating to the facial expression be determined and recorded, but also the raw three-dimensional data of a facial

scan and their corresponding photographic data. Through these measures, it is to be expected that the training results can be significantly improved and made more precise. Furthermore, cluster analysis and the introduction of more categories of facial expressions should be considered.

Currently, we are conducting the follow-up study with an optimized version of our app that encompasses the suggested alterations and adaptions. As we aim to build a large database of facial scans not only for facial expression recognition but also to monitor diseases' treatment progress and to perform disease recognition, we encourage fellow investigators to use our technology and contribute to our common database.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Conflict of Interest

The authors declare that they have no conflict of interest.

References

[1] Newmark C. Charles Darwin: The Expression of the Emotions in Man and Animals. Senge K, Schützeichel R (eds.). Hauptwerke der Emotionssoziologie. Wiesbaden: Springer Fachmedien Wiesbaden; 2013: 85–88. doi:10.1007/978-3-531-93439-6_11

[2] Barrett LF, Adolphs R, Marsella S et al. Emotional Expressions Reconsidered: Challenges to Inferring Emotion From Human Facial Movements. Psychol Sci Public Interest 2019; 20: 1–68. doi:10.1177/1529100619832930

[3] Kret ME, Maitner AT, Fischer AH. Interpreting Emotions From Women With Covered Faces: A Comparison Between a Middle Eastern and Western-European Sample. Front Psychol 2021; 12: 620632. doi:10.3389/fpsyg.2021.620632

[4] Tcherkassof A, Dupré D. The emotion-facial expression link: evidence from human and automatic expression recognition. Psychol Res 2021; 85: 2954–2969. doi:10.1007/s00426-020-01448-4

[5] Camerlink I, Coulange E, Farish M et al. Facial expression as a potential measure of both intent and emotion. Sci Rep 2018; 8: 17602. doi:10.1038/s41598-018-35905-3

[6] Burgoon JK. Microexpressions Are Not the Best Way to Catch a Liar. Front Psychol 2018; 9: 1672–1672. doi:10.3389/fpsyg.2018.01672

[7] ten Brinke L, Porter S, Baker A. Darwin the detective: Observable facial muscle contractions reveal emotional high-stakes lies. Evol Hum Behav 2012; 33: 411–416. doi:10.1016/j.evolhumbehav.2011.12.003

[8] Hurley CM, Frank MG. Executing Facial Control During Deception Situations. J Nonverbal Behav 2011; 35: 119–131. doi:10.1007/s10919-010-0102-1

[9] Porter S, ten Brinke L, Wallace B. Secrets and Lies: Involuntary Leakage in Deceptive Facial Expressions as a Function of Emotional Intensity. J Nonverbal Behav 2012; 36: 23–37. doi:10.1007/s10919-011-0120-7

[10] Haan J. Protagonists with Parkinson's disease. Front Neurol Neurosci 2013; 31: 178–187. doi:10.1159/000343237

[11] Girard JM, Cohn JF, Mahoor MH et al. Social Risk and Depression: Evidence from Manual and Automatic Facial Expression Analysis. Proc Int Conf Autom Face Gesture Recognit 2013. doi:10.1109/FG.2013.6553748

[12] Patel KR, Cherian J, Gohil K et al. Schizophrenia: overview and treatment options. P T 2014; 39: 638–645

[13] Foussias G, Remington G. Negative Symptoms in Schizophrenia: Avolition and Occam's Razor. Schizophr Bull 2008; 36: 359–369. doi:10.1093/schbul/sbn094

[14] Carbon C-C. Wearing Face Masks Strongly Confuses Counterparts in Reading Emotions. Front Psychol 2020; 11: 566886. doi:10.3389/fpsyg.2020.566886

[15] Sarkozy A, Digilio MC, Dallapiccola B. Leopard syndrome. Orphanet J Rare Dis 2008; 3: 13. doi:10.1186/1750-1172-3-13

[16] Ajitkumar A, Jamil RT, Mathai JK. Cri Du Chat Syndrome. StatPearls [Internet]. Treasure Island (FL): StatPearls Publishing; 2020

[17] Moramarco A, Himmelblau E, Miraglia E et al. Ocular manifestations in Gorlin-Goltz syndrome. Orphanet J Rare Dis 2019; 14: 218. doi:10.1186/s13023-019-1190-6

[18] John A, Abhishek MC, Ajayan AS, Sanoop S, Kumar VR. Real-Time Facial Emotion Recognition System With Improved Preprocessing and Feature Extraction. 2020 Third International Conference on Smart Systems and Inventive Technology (ICSSIT). Tamil Nadu, India: IEEE; 2020: 1328–1333. doi:10.1109/ICSSIT48917.2020.9214207

[19] Jeong M, Ko BC. Driver's Facial Expression Recognition in Real-Time for Safe Driving. Sensors (Basel) 2018; 18. doi:10.3390/s18124270

[20] Tian Y, Kanade T, Cohn JF. Facial Expression Recognition. Li SZ, Jain AK (eds.). Handbook of Face Recognition. London: Springer; 2011: 487–519. doi:10.1007/978-0-85729-932-1_19

[21] Lopes AT, de Aguiar E, De Souza AF et al. Facial expression recognition with Convolutional Neural Networks: Coping with few data and the training sample order. Pattern Recognit 2017; 61: 610–628. doi:10.1016/j.patcog.2016.07.026

[22] Huang Y, Chen F, Lv S et al. Facial Expression Recognition: A Survey. Symmetry 2019; 11: 1189. doi:10.3390/sym11101189

[23] Samadiani N, Huang G, Cai B et al. A Review on Automatic Facial Expression Recognition Systems Assisted by Multimodal Sensor Data. Sensors (Basel, Switzerland) 2019; 19: 1863. doi:10.3390/s19081863

[24] Chen Y, Du J, Liu Q et al. Robust and energy-efficient expression recognition based on improved deep ResNets. Biomed Tech (Berl) 2019; 64: 519–528. doi:10.1515/bmt-2018-0027

[25] Dawes TR, Eden-Green B, Rosten C et al. Objectively measuring pain using facial expression: is the technology finally ready? Pain Manag 2018; 8: 105–113. doi:10.2217/pmt-2017-0049

[26] Liu D, Cheng D, Houle TT et al. Machine learning methods for automatic pain assessment using facial expression information: Protocol for a systematic review and meta-analysis. Medicine (Baltimore) 2018; 97: e13421. doi:10.1097/MD.0000000000013421

[27] Bargshady G, Zhou X, Deo RC et al. Enhanced deep learning algorithm development to detect pain intensity from facial expression images. Expert Syst Appl 2020; 149: 113305. doi:10.1016/j.eswa.2020.113305

[28] Monwar MM, Rezaei S. Pain Recognition Using Artificial Neural Network. 2006 IEEE International Symposium on Signal Processing and Information Technology. Vancouver, BC, Canada: IEEE; 2006: 28–33. doi:10.1109/ISSPIT.2006.270764

[29] Haines N, Southward MW, Cheavens JS et al. Using computer-vision and machine learning to automate facial coding of positive and negative affect intensity. PLoS One 2019; 14: e0211735. doi:10.1371/journal.pone.0211735

[30] Lucey P, Cohn JF, Kanade T, Saragih J, Ambadar Z, Matthews I. The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression. 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition – Workshops. San Francisco, CA: IEEE; 2010: 94–101. doi:10.1109/CVPRW.2010.5543262

[31] Pantic M, Valstar M, Rademaker R, Maat L. Web-based database for facial expression analysis. 2005 IEEE International Conference on Multimedia and Expo. Amsterdam: IEEE; 2005. doi:10.1109/ICME.2005.1521424

[32] Valstar M, Pantic M. Induced disgust, happiness and surprise: an addition to the mmi facial expression database. Proc. 3rd Intern. Workshop on EMOTION (satellite of LREC): Corpora for Research on Emotion and Affect. London 2010: 65

[33] Susskind JM, Anderson AK, Hinton GE. The Toronto Face Database. Department of Computer Science, University of Toronto. Toronto, ON, Canada: Tech Rep; 2010: 3

[34] Goodfellow IJ, Erhan D, Carrier PL et al. Challenges in representation learning: A report on three machine learning contests. International conference on neural information processing. Daegu, South Korea: Springer; 2013: 117–124

[35] Dhall A, Murthy OVR, Goecke R, Joshi J, Gedeon T. Video and Image based Emotion Recognition Challenges in the Wild: EmotiW 2015. Proceedings of the 2015 ACM on International Conference on Multimodal Interaction. Seattle, Washington, USA: ACM; 2015

[36] Dhall A, Goecke R, Ghosh S, Joshi J, Hoey J, Gedeon T. From individual to group-level emotion recognition: EmotiW 5.0. ICMI '17: Proceedings of the 19 th ACM International Conference on Multimodal Interaction. New York: ACM; 2017: 524–528. doi:10.1145/3136755.3143004

[37] Gross R, Matthews I, Cohn J et al. Multi-PIE. Image Vision Comput 2010; 28: 807–813. doi:10.1016/j.imavis.2009.08.002

[38] Zhao G, Huang X, Taini M et al. Facial expression recognition from near-infrared videos. Image Vis Comput 2011; 29: 607–619. doi:10.1016/j.imavis.2011.07.002

[39] Zhang Z, Luo P, Loy CC et al. From facial expression recognition to interpersonal relation prediction. Int J Comput Vis 2018; 126: 550–569

[40] Mollahosseini A, Hasani B, Mahoor MH. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. IEEE Trans Affect Comput 2019; 10: 18–31. doi:10.1109/TAFFC.2017.2740923

[41] Li S, Deng W. Reliable Crowdsourcing and Deep Locality-Preserving Learning for Unconstrained Facial Expression Recognition. IEEE Trans Affect Comput 2019; 28: 356–370. doi:10.1109/TIP.2018.2868382

[42] Benitez-Quiroz CF, Srinivasan R, Martinez AM. EmotioNet: An Accurate, Real-Time Algorithm for the Automatic Annotation of a Million Facial Expressions in the Wild. 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2016: 5562–5570. doi:10.1109/CVPR.2016.600

[43] Goeleven E, De Raedt R, Leyman L et al. The Karolinska Directed Emotional Faces: A validation study. Cogn Emot 2008; 22: 1094–1118. doi:10.1080/02699930701626582

[44] Langner O, Dotsch R, Bijlstra G et al. Presentation and validation of the Radboud Faces Database. Cogn Emot 2010; 24: 1377–1388. doi:10.1080/02699930903485076

[45] Lijun Y, Xiaozhou W, Yi S, Wang J, Rosato MJ. A 3 D facial expression database for facial behavior research. 7 th International Conference on Automatic Face and Gesture Recognition (FGR06). Southampton, UK: University of Southampton; 2006: 211–216. doi:10.1109/FGR.2006.6

[46] Habibu R, Syamsiah M, Hamiruce MM, Iqbal SM. UPM-3 D Facial Expression Recognition Database (UPM-3DFE). PRICAI 2012: Trends in Artificial Intelligence. Berlin, Heidelberg: Springer; 2012: 470–479

[47] Cao C, Weng Y, Zhou S et al. FaceWarehouse: a 3 D facial expression database for visual computing. IEEE Trans Vis Comput Graph 2014; 20: 413–425. doi:10.1109/TVCG.2013.249

[48] Ertugrul IO, Cohn JF, Jeni LA et al. Cross-domain AU Detection: Domains, Learning Approaches, and Measures. Proc Int Conf Autom Face Gesture Recognit 2019; 2019: 1–8. doi:10.1109/FG.2019.8756543

[49] Yin L, Chen X, Sun Y, Worm T, Reale M. A high-resolution 3 D dynamic facial expression database. 2008 8 th IEEE International Conference on Automatic Face & Gesture Recognition. Amsterdam: IEEE; 2008: 1–6. doi:10.1109/AFGR.2008.4813324

[50] Liu YC, Kuo RL, Shih SR. COVID-19: The first documented coronavirus pandemic in history. Biomed J 2020; 43: 328–333. doi:10.1016/j.bj.2020.04.007

[51] United Nations Sustainable Development Group. Policy Brief: Education during COVID-19 and beyond. 2020. . Accessed September 01, 2020 at: https://unsdg.un.org/resources/policy-brief-education-during-covid-19-and-beyond

[52] Schleicher A. The impact of COVID-19 on Education – Insights from Education at a Glance 2020. 2020 . Accessed December 22, 2020 at: https://www.oecd.org/education/the-impact-of-covid-19-on-education-insights-education-at-a-glance-2020.pdf

[53] Tuttle KR. Impact of the COVID-19 pandemic on clinical research. Nat Rev Nephrol 2020; 16: 562–564. doi:10.1038/s41581-020-00336-9

[54] Daroedono E, Erwin F, Alfarabi M et al. The impact of COVID-19 on medical education: our students perception on the practice of long distance learning. Int J Community Med Public Health 2020. doi:10.18203/2394-6040.ijcmph20202545

[55] Di Pietro G, Biagi F, Dinis Mota Da Costa P, Karpinski Z, Mazza J. The likely impact of COVID-19 on education: Reflections based on the existing literature and recent international datasets. Luxemburg: Publications Office of the European Union; 2020.

[56] Joint Review Commission on Education in Diagnostic Medical Sonography (JRC-DMS). JRC-DMS Covid-19 Statement. 2020 . Accessed September 05, 2020 at: https://www.jrcdms.org

[57] Society for Vascular Ultrasound. Vascular Laboratory Responses During the COVID-19 Pandemic. 2020 . Accessed September 05, 2020 at: https://www.svu.org/svu-news/4183/

[58] Hilburg R, Patel N, Ambruso S et al. Medical Education During the Coronavirus Disease-2019 Pandemic: Learning From a Distance. Adv Chronic Kidney Dis 2020; 27: 412–417. doi:10.1053/j.ackd.2020.05.017

[59] Alsoufi A, Alsuyihili A, Msherghi A et al. Impact of the COVID-19 pandemic on medical education: Medical students' knowledge, attitudes, and practices regarding electronic learning. PLoS One 2020; 15: e0242905. doi:10.1371/journal.pone.0242905

[60] Bundesamt für Gesundheit (BAG). Coronavirus: Massnahmen und Verordnungen. 2020 . Accessed December 19, 2020 at: https://www.bag.admin.ch/bag/de/home/krankheiten/ausbrueche-epidemien-pandemien/aktuelle-ausbrueche-epidemien/novel-cov/massnahmen-des-bundes.html

[61] Hartmann T, Friebe-Hoffmann U, Gregorio N et al. Novel and flexible ultrasound simulation with smartphones and tablets in fetal echocardiography. Arch Gynecol Obstet 2022; 305: 19–29. doi:10.1007/s00404-021-06102-x

[62] Hartmann T, Friebe-Hoffmann U, Polasik A et al. Fetale Echokardiographie via Scanbooster Ultraschall Simulator App üben – wie verhält sich diese neue Lernmethode in Bezug auf Effektivität und Motivation Studierender? Geburtshilfe Frauenheilkd 2020; 80: P099

[63] Hartmann T, Friebe-Hoffmann U, Polasik A et al. Scanbooster Ultraschall Simulation mit Smartphone und Tablet in der Geburtshilfe. Geburtshilfe Frauenheilkd 2020; 80: P098

[64] Hartmann TJ, Friebe-Hoffmann U, Lato C et al. VP34.17: Practicing fetal echocardiography with the Scanbooster ultrasound simulator app on smartphone and tablet. Ultrasound Obstet Gynecol 2020; 56: 200–201. doi:10.1002/uog.22850

[65] Hartmann TJ, Friebe-Hoffmann U, Lato C et al. OC10.08: Comparing a new form of ultrasound simulation on smartphone and tablet to a conventional learning method. Ultrasound Obstet Gynecol 2020; 56: 30. doi:10.1002/uog.22272

[66] Forchhammer S, Hartmann T. Digitale Dermatopathologie: Vorteile für Befundung, Forschung und Ausbildung. Der Deutsche Dermatologe 2021; 69: 810–813. doi:10.1007/s15011-021-4760-6

[67] Chipps J, Brysiewicz P, Mars M. A Systematic Review of the Effectiveness of Videoconference-Based Tele-Education for Medical and Nursing Education. Worldviews Evid Based Nurs 2012; 9: 78–87. doi:10.1111/j.1741-6787.2012.00241.x

[68] Shadat A, Sayem M, Taylor B et al. Effective use of Zoom technology and instructional videos to improve engagement and success of distance students in Engineering. Australas J Eng Educ 2017; 22: 926–931

[69] Sathik M, Jonathan GS. Effect of facial expressions on student's comprehension recognition in virtual educational environments. Springerplus 2013; 2: 455. doi:10.1186/2193-1801-2-455

[70] Apple Inc.. Informationen zur fortschrittlichen Technologie von Face ID. 2020 . Accessed July 23, 2020 at: https://support.apple.com/de-de/HT208108

[71] Apple Inc.. ARFaceAnchor | Apple Developer Documentation. 2020 . Accessed August 03, 2020 at: https://developer.apple.com/documentation/arkit/arfaceanchor

[72] Apple Inc.. ARKit | Apple Developer Documentation. 2020 . Accessed August 03, 2020 at: https://developer.apple.com/documentation/arkit

[73] Apple Inc.. geometry | Apple Developer Documentation. 2021 . Accessed January 15, 2021 at: https://developer.apple.com/documentation/arkit/arfaceanchor/2928271-geometry

[74] Apple Inc.. ARFaceAnchor.BlendShapeLocation | Apple Developer Documentation. 2020 . Accessed August 03, 2020 at: https://developer.apple.com/documentation/arkit/arfaceanchor/blendshapelocation

[75] Apple Inc.. blendShapes | Apple Developer Documentation. 2020 . Accessed August 03, 2020 at: https://developer.apple.com/documentation/arkit/arfaceanchor/2928251-blendshapes

[76] Apple Inc.. Dictionary | Apple Developer Documentation. 2021 . Accessed January 15, 2021 at: https://developer.apple.com/documentation/swift/dictionary

[77] Apple Inc.. shuffle() | Apple Developer Documentation. 2020 . Accessed July 24, 2020 at: https://developer.apple.com/documentation/swift/array/2994753-shuffle

[78] Abadi M, Agarwal A, Barham P et al. Tensorflow: Large-scale machine learning on heterogeneous distributed systems. Proceedings of the 12 th USENIX conference on Operating Systems Design and Implementation (OSDI'16. Berkeley: USENIX; 2016: 265–283

[79] Srivastava N, Hinton G, Krizhevsky A et al. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. J Mach Learn Res 2014; 15: 1929–1958

[80] Google Ireland Limited. tf.keras.utils.plot_model : TensorFlow Core v2.3.0. 2020 . Accessed August 04, 2020 at: https://www.tensorflow.org/api_docs/python/tf/keras/utils/plot_model

[81] Mutasa S, Sun S, Ha R. Understanding artificial intelligence based radiology studies: What is overfitting? Clin Imaging 2020; 65: 96–99. doi:10.1016/j.clinimag.2020.04.025

[82] Klambauer G, Unterthiner T, Mayr A et al. Self-Normalizing Neural Networks. arXiv 2017. doi:10.48550/arXiv.1706.02515

[83] Kingma D, Ba J. Adam. A method for stochastic optimization. arXiv 2017. doi:10.48550/arXiv.1412.6980

[84] Dodge Y, Institute IS, Commenges D. The Oxford Dictionary of Statistical Terms. Oxford; UK: Oxford University Press; 2006.

[85] Google Ireland Limited. tf.keras.activations.selu | TensorFlow Core v2.3.0. 2020 . Accessed August 04, 2020 at: https://www.tensorflow.org/api_docs/python/tf/keras/activations/selu?hl=en

[86] LeCun Y, Bottou L, Orr G, Müller K-R. Efficient BackProp. Montavon G, Orr GB, Müller K-R (eds.). Neural Networks: Tricks of the Trade Lecture Notes in Computer Science. Berlin, Heidelberg: Springer; 2012: 9–48. doi:10.1007/978-3-642-35289-8_3

[87] Google Ireland Limited. tf.keras.layers.AlphaDropout | TensorFlow Core v2.3.0. 2020 . Accessed August 04, 2020 at: https://www.tensorflow.org/api_docs/python/tf/keras/layers/AlphaDropout?hl=en

[88] Cortes C, Mohri M, Rostamizadeh A. L2 regularization for learning kernels. arXiv 2012. doi:10.48550/arXiv.1205.2653

[89] Anguita D, Ghelardoni L, Ghio A, Oneto L, Ridella S. The 'K' in K-fold Cross Validation. ESANN 2012 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 25–27 April 2012. Louvain-la-Neuve, Belgium: i6doc.com; 2012

[90] Breitbarth A, Schardt T, Kind C, Brinkmann J, Dittrich P-G, Notni G. Measurement accuracy and dependence on external influences of the iPhone X TrueDepth sensor. Proc. SPIE 11144, Photonics and Education in Measurement Science 2019, 1114407 (17 September 2019) 2019. doi:10.1117/12.2530544