Hindawi
Journal of Electrical and Computer Engineering
Volume 2023, Article ID 9351345, 11 pages
https://doi.org/10.1155/2023/9351345

Research Article

Designing an Efficient System for Emotion Recognition Using CNN

Donia Ammous,[1,2] **Achraf Chabbouh**,[3,4] **Awatef Edhib**,[2] **Ahmed Chaari**,[4]
Fahmi Kammoun,[1] **and Nouri Masmoudi**[1]

[1]*National School of Engineers of Sfax, University of Sfax, Laboratory of Electronics and Information Technologies, Circuit and System Team "C&S", LR99ES37, 3038 Sfax, Tunisia*
[2]*Sogimel: A Consulting Company in Computer Engineering and Video Surveillance, Sfax Technopole, 3021 Sakiet Ezzit, Tunisia*
[3]*Higher Institute of Technological Studies of Sidi Bouzid, Sidi Bouzid, Tunisia*
[4]*Anavid France, 10 Road Penthièvre, Paris, France*

Correspondence should be addressed to Donia Ammous; ammous.donia@gmail.com

Received 4 February 2023; Revised 18 August 2023; Accepted 23 August 2023; Published 15 September 2023

Academic Editor: Yang Li

Implementing an efficient system for emotion recognition has recently posed a challenge that has not been fully developed yet. Facial emotion recognition (FER) is an important subject matter in the fields of artificial intelligence (AI) since it exhibits a greater commercial potential. This technique is used to analyse various sentiments and reveal a person's behavior. It could be related to the mental or physiological state of mind. This paper mainly focuses on a human emotion recognition system through a detected human face. Its accuracy was improved via different data augmentation tools, early stopping, and generative adversarial networks (GANs). Compared to previous methods, experimental results show that the proposed method provides a 0.55% to 35.7% gain performance.

1. Introduction

Various techniques have emerged to perform emotion recognition from faces including support vector machines [1], hidden Markov models [2], and neural networks [3, 4]. In machine learning, SVM is applied as categorizing and reverting exploration. It is a supervised machine learning algorithm that can be used in resolving data classification and regression challenges.

SVM was developed and created from statistical learning theory by Vapnik and Chervonenkis in 1990. Its basic idea is to transform the input space to a high-dimensional feature space using a kernel function and then obtain a maximum classification in this new feature space.

Hidden Markov models (HMMs) have been widely used to model the temporal behaviors of facial expressions from image sequences. HMMs are able to model temporal dependencies. HMMs are probabilistic models which consist of a countable number of states, transitions, and corresponding emissions. HMMs are easy to model but variable by the parameters that describe them.

Researchers proposed many techniques based on neural network's approach to recognize the facial expressions. The neural networks can easily implement the mapping from the feature space of face images to the facial expression space.

In addition, neural networks can work well for tackling the pattern classification issues in engineering [5, 6]. The neural network has been used in several problems in image classification for image processing. Researchers have developed various NN structures in accordance with their problem. After the network is trained, the generative model can be tested. To achieve optimization, you have the flexibility to fine-tune hyperparameters and customize the neural network architecture by incorporating specific layers. In fact, a variety of optimization tools has been employed in neural networks to learn from past experiences and use that prior training to identify new patterns and classify new sentimental data.

We have recently witnessed a progress in the use of deep learning (DL) for neural networks leading to classification [7–10]. Facial expression recognition in videos is an active area of research in computer vision. Accordingly, researchers in this field are interested in developing techniques to interpret facial expressions, code them, and extract their features in order to have a better prediction of emotions. Capitalizing on the remarkable success of deep learning, various architecture types are harnessed to attain superior performance. The remainder of this paper is organized as follows: Section 2 reviews and discusses the previous approaches used for expression recognition. Section 3 describes the proposed work. Section 4 introduces the results and discussion. Section 5 concludes this paper.

2. Previous Work

Sinha and Aneesh [9] embedded the handcrafted features in the training process of the network in order to reduce the difference between the features learned by deep networks and the handcrafted ones. Their work was based on the HoloNet with feature loss (HNwFL) for feature learning and the fusion network for recognition. In HNwFL, the handcrafted feature information was integrated into the HoloNet and the new feature loss was tested on CK+, JAFFE, and FER2013 datasets. When compared with other works, this network obtained the best accuracy with a 97.35% on the CK+ dataset. In fact, the suggested network provided much better accuracy than the network which did not include feature loss as well as the original handcrafted feature.

In the study in [11], the initial work of both computer vision and image processing has been developed in order to facilitate the teaching of young autistic children recognizing the human facial expression. Facial expression recognition work was proposed by Haque et al. using a deep convolutional neural network. In order to experiment and train the deep convolutional neural network model, Kaggle's FER2013 dataset was used. Their work resulted in 63.11% accuracy without overfitting the model. In fact, the bright images achieved a better accuracy than the dark ones. Thus, the dataset was modified across four groups with different lighting conditions, and each set is again trained with the same model. Mollahosseini et al. [12] advanced deep neural network architecture to address the FER problem via well-known face datasets. Seven publically available databases, mainly MultiPIE, MMI, CK+, DISFA, FERA, SFEW, and FER2013, were used in this experiment. The proposed network consisted of two convolutional layers. Each layer was followed by max pooling and then by four inception layers. The architecture network took facial images as an input and classified them into six emotional expressions (anger, disgust, fear, happiness, sadness, and surprise). Its results were comparable to or more efficient than traditional convolutional neural networks in both accuracy and training time. Tümen et al. [13] aimed to build a convolutional neural network (CNN) based on a facial expression recognition (FER) system so as to classify expressions presented in the FER2013 database. The presented CNN achieved a 57.1% accuracy rate on the FER2013 database [13]. It also provided a good result while detecting emotional expressions without any preprocessing. A high rate was particularly achieved in three classes of happiness, surprise, and disgust. Xie et al. [14] introduced a novel approach for FER named deep attentive multipath convolutional neural networks (DAM-CNNs). The proposed model contained three modules including the VGG-Face, the Salient Expressional Region Descriptor (SERD), and the Multipath Variation-Suppressing Network (MPVS-Net). The VGG-Face (visual geometry group) extracted a feature. The SERD automatically located expression-related regions in the target image. The MPVS-Net module separated expressional information from irrelevant variations. By jointly combining SERD and MPVS-Net, the DAM-CNN was able to highlight relevant facial traits and yield a robust image representation for emotion recognition. AlMarri [15] used a fast region-based convolutional neural network (Fast R-CNN) for facial emotion recognition. The proposed approach aimed at training several network models with different training options. Several experimental tests were conducted to validate both network convergence and generalization. The network models, which were trained on the Ryerson Audio-Visual Database of Emotional Speech and Song (RAVDESS) library, reached the best result. The network model could classify facially expressed emotions for people of different ages and ethnicities such as happy and surprised, achieving 100% and 94%, respectively.

3. Background

The current paper is based on the studies of Github et al.'s [16, 17] technique which was selected according to different tests presented in the "Study of previous works" section. This method was modified to boost the performance of its neural network. Github et al.'s [16, 17] approach is presented through the neural network architecture summarized in Figure 1. Initially, a lot of images were captured with a camera. The face extraction module then used trained Haar cascade/deep neural networks (DNNs). The classification algorithm was trained on the FER2013 dataset. Finally, a model was generated to classify emotions through different facial expressions such as angry, disgusted, fearful, happy, sad, surprised, and neutral.

3.1. Proposed Method. The facial expression recognition was difficult to realize due to the slight difference between several emotions, which required an efficient and precise algorithm to be trained.

The imbalanced distributions of emotion classes gave rise to low accuracy. Therefore, two fundamental rules, notably the data-centric machine learning (ML) and the model-centric approach [18], were used in deep learning algorithms. The data-centric ML aimed at improving the quality of the used dataset by

(i) Balancing classes: the number of items should be the same in each class

(ii) Data augmentation: adding augmented images to some classes

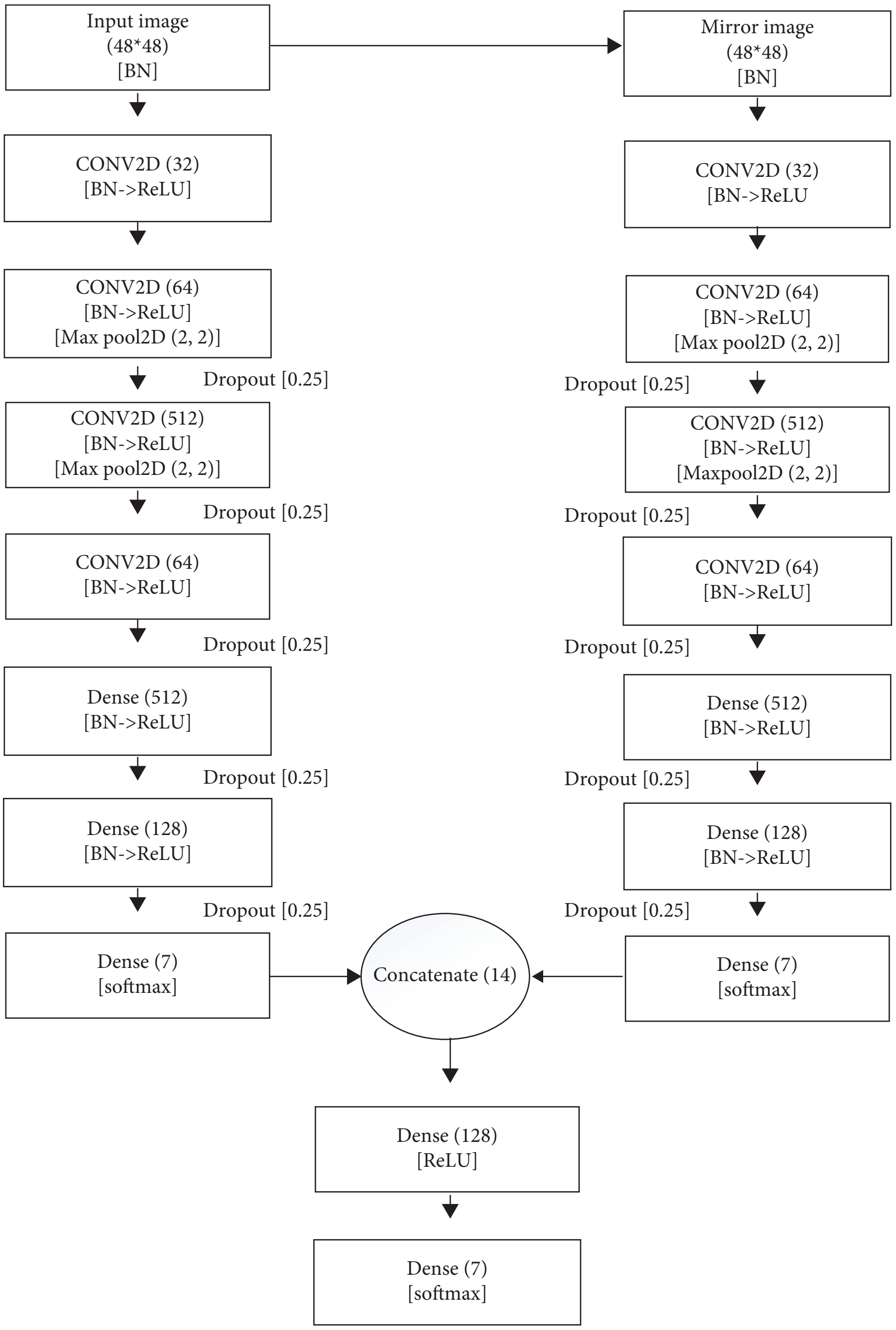


Figure 1: Neural network architecture.

(iii) Increasing diversity: images should be from shops (high quality) and consumer-taken (low quality)

(iv) Removing classes

The model-centric approach included two techniques:

(i) Early stopping method using Keras

(ii) Hyperparameter tuning

Two enhancement strategies were adopted to overcome the shortcomings of the Kant et al. method and improve its performance. The first strategy accentuated the RelGAN (relative attributes generative adversarial networks) approach [19, 20], while the second introduced the early stopping technique. These two strategies were explored in the functional block diagram of our implementation, as shown in Figure 2. First, the anavid dataset [21] was collected from both private and public datasets and its preparation was optimized in three versions. Then, various data augmentation techniques were exercised on the anavid dataset. Moreover, the RelGAN method was performed on the same dataset. After creating our own dataset, the resulted images were divided into train, test, and validation sets as per its

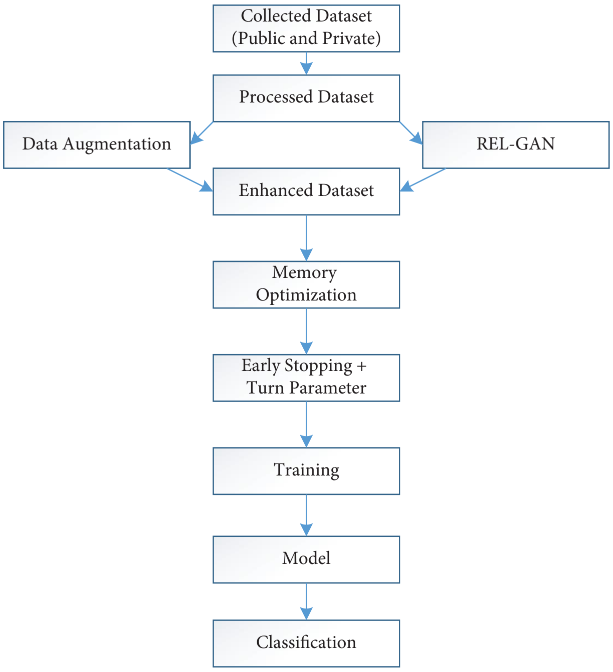


Figure 2: Block diagram of proposed method.

usage. Afterwards, memory optimization was integrated into the algorithm. Subsequently, early stopping was applied to the neural network architecture. Besides, the designed model was ameliorated by tuning some potential hyperparameter combinations. Finally, the emotions' classification was provided.

Three contributions were described in detail in three bullets, and in each bullet, it explains one contribution clearly (data augmentation, generative adversarial networks (GAN), and early stopping).

3.1.1. Data Augmentation. Data augmentation is a widely employed method in computer vision, which entails applying diverse transformations to the initial training data, thereby generating additional augmented samples. By adopting this approach, the size and diversity of the training dataset can be expanded, leading to enhanced performance and robustness of computer vision models. Both input images and the labels for those images can be enhanced with data using certain techniques. In computer vision, the following are some typical data augmentation techniques:

(i) Flipping or mirroring: This entails flipping the image either horizontally or vertically to produce a mirrored replica. It aids in improving the model's ability to generalize to various orientations or viewpoints.

(ii) Rotation: Rotating the image by a predetermined angle, such as 90 degrees or any other predetermined angle, might assist models become invariant to rotation fluctuations.

(iii) Scaling and cropping: Rescaling the image to multiple sizes or randomly clipping a section of the image helps increase the model's robustness and ability to handle objects of various scales.

(iv) Translation: Shifting the image horizontally or vertically makes models more tolerant of object displacement and helps them learn spatial invariance.

(v) Gaussian noise: Incorporating random noise into the image simulates variations in lighting conditions and enhances the model's robustness in noisy environments.

(vi) Color jittering: Changing the image's color values, such as brightness, contrast, or saturation, might help the model better adapt to changes in the color and lighting of an environment.

These are just a few instances of computer vision techniques for data augmentation. The selection and combination of the augmentation methods depend on the particular task, dataset, and the desired characteristics that the model should learn. Data augmentation is a useful technique for enhancing computer vision models' performance and generalization, particularly when there is a lack of training data.

3.1.2. Generative Adversarial Network (GAN). The generative adversarial network (GAN), defined as a powerful unsupervised generative model, was widely applied in order to generate photorealistic images and a variety of other image types. It is, thus, beneficial to classify emotion recognition and train data augmentation. An analysis study of the GAN method was performed leading to a whole GAN family identification including CycleGAN, DCGAN, MoCoGAN, and RelGAN [22–24] which are representative methods in a multidomain image-to-image translation. The RelGAN technique was selected, and its code was modified. The RelGAN technique consists of generating the same face with other expressions and looks ("Pale_Skin," "Smiling," "Eyeglasses," and "Gray_Hair"). This method is capable of modifying images by changing particular attributes of interest [19, 20]. Four attributes were applied to public datasets:

(i) Pale skin

(ii) Smiling

(iii) Eyeglasses

(iv) Gray_Hair

Figure 3 illustrates the original image and its four effects ("pale skin," "Smiling," "Eyeglasses," and "Gray_hair"). Tests revealed the efficiency of the RelGAN technique on a public dataset because the public images were taken from a close distance and the facial features were therefore clear. Using this technology, a DL algorithm could generate more facial expressions. Examples of the RelGAN technology are presented in Figure 3.

3.1.3. Early Stopping. Early stopping [25, 26] is an effective technique which is applied to prevent the overfitting

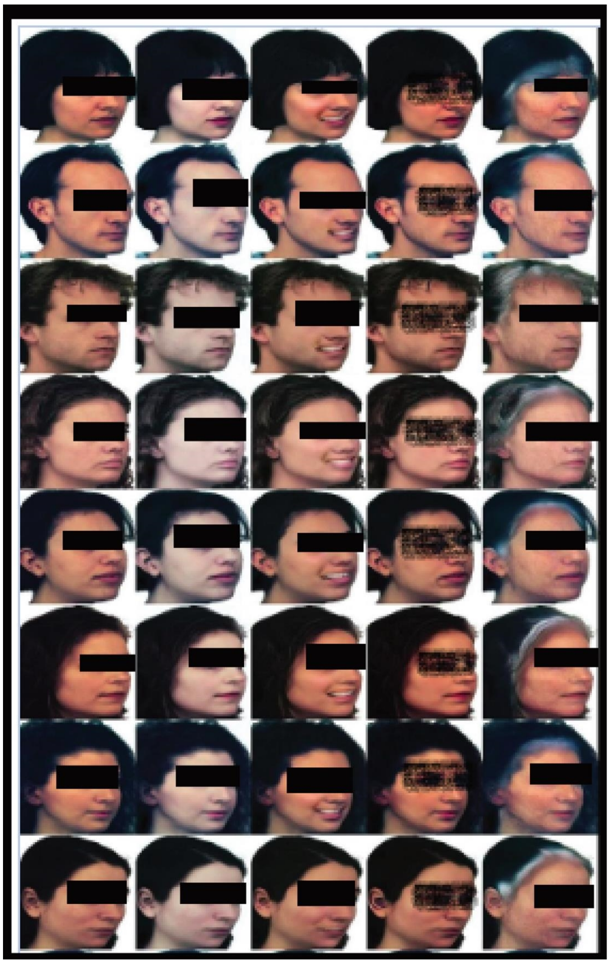


Figure 3: RelGAN examples.

phenomena and improve the generalization of deep neural networks. It is a form of regularization achieved through choosing an arbitrary number of training epochs and stopping the training process early before it overfits its dataset. The training of a neural network will be arrested by early stopping once the model performance stops and the accuracy ceases progressing on the validation set. During training, the model with the best validation dataset accuracy is selected. The early stopping algorithm is a technique used for reducing computational algorithms. In experimental results, different hyperparameters such as monitored quantities and patience are tuned. A small patience will produce an undertrained model, whereas a big one will lead to an overtrained model. If a monitor stops decreasing, the training of the model will stop. The monitor can be training loss and validation loss. The patience is the number of epochs with no decrease in the monitor. In order to regularize the machine learning model, early stopping stops training when the validation loss reaches the minimum.

The experiments show that early stopping can enable the networks to learn more novel features and have a high predictive performance.

4. Experimental Results

4.1. Datasets. A series of experiments was performed over different datasets:

JAFFE [27]: The Japanese Female Facial Expression (JAFFE) database contains 213 images of posed expressions from 10 Japanese female subjects. Each subject represents 7 different emotional facial expressions (6 basic facial expressions + 1 neutral). The database is challenging since it consists of a little example images per subject/expression. This dataset is asked to do acted facial expressions related to Eckman's emotions, including the following facial expressions: "happiness," "anger," "sadness," "surprise," "disgust," "fear," and "neutral." The resolution of original facial images is 256 × 256 pixels with tiff format. Several dataset images of each expression are assembled by Miyuki Kamachi, Michael Lyons, and Jiro Gyoba, at Kyushu University, Japan. Each image is annotated with average semantic ratings on nouns describing the posed expression by 60 Japanese viewers.

Facial expression recognition (FER) 2013 [28]: The FER2013 database was presented during the ICML 2013. The dataset was labeled and created by the Google search API. It [29] is used in 32% research studies and consists of 35,887 images. In fact, FER2013 includes 28,709 training sets, 3,589 validation images, and 3,589 test sets. In order to label all the 36k photos, the dataset uses Ekman's emotions which are composed by seven emotional expressions (neutral, fear, disgust, sadness, happiness, surprise, and anger). Each image is in grayscale with a resolution of 48 × 48 pixels. The images are smaller than other datasets. FER has more variation in the frames, including facial occlusion (mostly with a hand), partial faces, low-contrast images, and eyeglasses.

KDEF [30]: Karolinska Directed Emotional Faces (KDEF) comprised images from 70 individuals: 35 men and 35 women, all between 20 and 30 years old, where each image belonged to one of the following classes (fear, anger, disgust, happiness, neutrality, sadness, and surprise) and each expression is photographed from five different angles (full left profile, full right profile, half left profile, right half profile, and straight profile). The KDEF dataset represents color photographs with a 562 × 762 image format. The KDEF dataset [31] consisted of 4,900 images which were made at the Karolinska Institute, Stockholm. It is widely used in the field of facial expression recognition.

The extended Cohn–Kanade (CK+) [32] is a popular facial expression recognition dataset and is commonly used in several works. The CK+ dataset contains 593 image sequences of persons with different age (18 to 50 years old), gender, and heritage. The dataset contains images with a resolution of 640 × 490. Among all, 327 sequences are labeled with an emotion in seven basic expression labels: anger, contempt, disgust, fear, happiness, sadness, and surprise.

(MMA Facial Expression Database MMAFEDB) [32] is collected from different emotion and expression images with a resolution of 48 × 48 pixels. It is divided into 3 sections as testing, training, and validating. Each section includes seven facial expression categories: angry, disgusted, fearful, happy, neutral, sad, and surprised. These three sections contain 17,356 testing, 92,968 training, and 17,356 validating.

The Yale face database [33] contains 165 grayscale images of 15 individuals with different facial expressions including center-light, w/glasses, happy, left-light, w/no glasses, normal, right-light, sad, sleepy, surprised, and wink.

The GENKI database [34] was collected by the Machine Perception Laboratory (MPLab), University of California, San Diego. The MPLab GENKI-4K is a database subset containing 4,000 real-life faces annotated in two classes of images "smiling" and "nonsmiling," downloaded from publicly available Internet repositories. These images involve a wide range of backgrounds, geographical locations, illumination conditions, personal identity, and ethnicity. This dataset has large variations in pose, age, and gender. It consists of 1,940 grayscale face images with varying facial expressions. The GENKI-SZSL subset includes 3,500 images collected from the Internet. They are classified according to face location and size. The current release of the GENKI dataset is GENKI-R2009a, a version which consists of 7,172 images that combine to form the following subsets:

(1) GENKI-4K: 4,000 images representing expression and head-pose labels

(2) GENKI-SZSL: 3,500 images introducing face position and size labels

A training set consisting of 2,958 data samples is selected from the whole GENKI dataset.

4.2. Evaluation Criteria. Confusion matrix [35, 36], an evaluation measure broadly applied in deep learning classification algorithms, is used to reflect the behavior of various models in supervised classification contexts. It is a square matrix in which the rows represent the instances' actual class and the columns show their predicted class. The following 2 × 2 confusion matrix reports the number of true positives (TP), false negatives (FN), false positives (FP), and true negatives (TN).

(i) True negatives are the number of negative elements that are correctly predicted as negatives

(ii) True positives are the number of positive elements that are correctly labeled as positives

(iii) False negatives are the number of positive elements that are classified falsely as negatives

(iv) False positives are the number of negative elements that are incorrectly classified as positives

4.3. Tests Conditions and Improved Source Materials. Our tests were realized in Jetson AGX Xavier. Thanks to NVIDIA Jetson AGX Xavier [36], AI robotics applications were easily deployed and created for manufacturing, retail, delivery, agriculture, and more. Jetson Xavier NVIDIA is an AI computer for autonomous machines offering a GPU (graphical processing unit) performance of less than 30 W supported by NVIDIA JetPack 4.2. Jetson AGX Xavier brand NVIDIA and graphic coprocessor NVIDIA Volta and CUDA 10 with memory LPDDR4X 256 bits/16 Go.

Memory management [37] is the biggest challenge in deep neural networks (DNNs). In fact, memory can still be an important constraint when the model size is too big, working with large amounts of sets, or running two models at the same time.

During training, the use and the generation of the Keras models require huge memory. Thus, the optimization for embedded GPUs is treated to fit in with the limited memory bandwidth and potentially decrease memory storage costs.

In the proposed approach, we intend to reduce memory computation. In fact, the integration of some instructions is applied in the CNN algorithm to realize real-time processing. The following instructions are added:

```
import tensorflow as tf
from Keras import backend as k
gpu_options = tf.GPUOptions
(per_process_gpu_memory_fraction = 0.1)
sess = tf.Session
(config = tf.ConfigProto(gpu_options = gpu_options))
```

The memory consumption optimization of facial emotion recognition is obtained using the convolution neural network (CNN). In fact, optimization memory is integrated into the algorithm with a variable parameter (0.192 and 0.1). As illustrated in Table 1, the proposed experiment reduces memory consumption from 9.07 to 4.89. Hardware resource utilization is also reduced from 4.89 to 1.64 by turning the parameter to 0.1 (see Table 1). As a result, the current approach offers a flexible trade-off between the high accuracy of the overall adopted network and the lower memory consumption.

4.4. Study of Previous Works. In this section, three previous methods conducted by Github et al. [16, 17], Li et al. [38], and Correa et al. [39] are compared. The Github et al. method [16, 17] is described in the abovementioned section of background. Li et al. [38] proposed a multikernel convolutional block to extract facial expression features. First, this approach was designed through three depth-wise separable convolutional kernels. Second, the multichannel information was fused to obtain multikernel enhancement features. Then, a "channel split" task was performed on the multikernel convolutional block input. Finally, a lightweight multikernel feature expression recognition network was designed by alternately using the multikernel convolutional block and the depth-wise separable convolutions. Correa et al. [39] designed a neural network based on an artificial intelligent system for emotion recognition. Three promising neural network architectures were trained and subjected to various facial expressions' classification tasks. Then, the best performing network was further improved. For this purpose, different approaches were experimented and evaluated. The final model was portrayed and applied in a real video stream application that could instantaneously return the user emotion. In this paper, deeply trained models for emotion

Table 1: Memory consumption optimization (gigabyte).

With optimization "0.192"	With optimization "0.1"	Without optimization
4.89	1.64	9.07

detection are presented through the use of the FER emotion databases. The Github et al. method [16, 17] achieved the best results compared to the Correa et al. [39] and Li et al. [38] methods. The simulation has the most prominent values in different classes (angry, happy, sad, and neutral). This highly efficient model of the Kant et al. method is explained by high true positive values in four categories (see Table 2). Consequently, the Kant et al. method is used in all remainder tests. In order to further ameliorate its performance, different tools such as data augmentation, RelGAN technology, and early stopping are performed.

4.5. Model Comparison (with and without Data Augmentation). The small dataset affects the trained model. They do not generate reliable data from the test and validation image. Thus, these generated models suffer from the overfitting issue which can be solved through various proposed methods. According to the study in [40], the principle of this approach is to add regularization to the weights. Another technique is a dropout technique [41]. This principle consists of dropping certain connections in layers or removing a neuron from layers during training. Another popular method is batch normalization [42]. It is applied to any neural network layer. Some works explore data augmentation techniques to resolve the overfitting problem. In deep convolutional neural networks, the limited numbers of pictures contribute to the overfitting phenomena. During training, the number of samples in each class is too small and the rate of accuracy is wrong. For this reason, a different kind of data augmentation is applied to the dataset. Various data augmentation techniques such as rotating, noise adding, flipping, zoom, cropping, and brightness are performed. As a result, the loss classification is reduced, and the neural network learns better and generates the best model. This strategy is an effective way to improve the accuracy of image classifiers. As it can be noticed in Table 3, data augmentation is an effective method for image classification. It can be observed that the balanced dataset model based on data augmentation achieves a high true positive value of four emotion categories compared to unbalanced dataset models.

The most facial landmarks are the mouth, the corners of the eyes, the lobes of the ears, the chin, and the tip of the nose. This helps to distinguish emotion class for pictures captured in the front of the camera. In the case of blurry images or pictures captured by the position away from the camera, the algorithm provides incorrect classification (see Table 4).

This model can be improved by adding a bunch of datasets and collecting enough data for training using the RelGAN technique. In the following section, future experimentations may be fulfilled using various datasets.

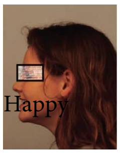
Table 2: Class distribution.

Method/FER dataset	True positives			
	Angry	Happy	Sad	Neutral
Method of Enrique et al. [39]	396	1,507	395	530
Method of Minze et al. [38]	406	790	515	502
Method of Kant et al. [16, 17]	519	1,524	790	636

Table 3: Data augmentation effect.

	True positives			
	Angry	Neutral	Happy	Sad
Unbalanced dataset model	1,289	1,753	1,228	1,122
Balanced dataset model	1,362	1,789	1,746	1,305

Table 4: Example of images.

Emotion of ground truth	Prediction of emotion
	Angry
	Happy
	Angry
	Angry

4.6. Comparative Study of a Combination of Early Stopping and Data Augmentation Test. The early stopping can be combined with data augmentation in order to enhance the model. Early stopping is added to prevent deep learning neural network models from overfitting phenomena. The early stopping technique provides three benefits including the following:

(i) Reduce the overfitting phenomena by adding early stopping to an existing model

(ii) Reduce the training execution time by choosing the number of training epochs

(iii) High rate in the confusion matrix

Table 5: Parameter setting.

Parameter	Value
Mode	Min
Patience	3
Min_delta	0.00001
Monitor	Val_loss
Verbose	1
Restore_best_weights	True

The early stopping method is realized in the implementation phase. The hyperparameters of early stopping are adjusted to further improve the model. The hyperparameters are summarized in Table 5.

In this section, specific parameters are described, including mode, monitor, patience, verbose, min_delta, and restore_best_weights.

"mode": mode is set to min. We seek the minimum for validation loss and the maximum for validation accuracy.

"monitor": Quantity to be monitored.

"patience": The number of epochs without decreasing the monitor after the training stops. For example, if patience is set to 3 without decreasing the loss for three consecutive epochs, the training stops.

The "verbose" can be tuned to 1. Once the training is stopped, the epoch number is printed.

"min_delta": min_delta is fixed to 0.00001. Therefore, an absolute change of less than min_delta cannot be considered as an improvement.

"restore_best_weights": The restored best weight is fixed to true positives to make sure the final model we get is the best.

Supervised machine learning (ML) algorithms require a lot of data to be able to generalize models. The principle of data augmentation is to generate more images from the original dataset and create more variations of image appearance such as backgrounds and different contexts. Therefore, during training, more image variations are obtained to boost the performance of the recognition model. In order to increase both data size and training data diversity, various transformations are used.

As a result, the dataset anavid [21] is modified in three different sets and each of these datasets is again trained. Table 6 contains these datasets with distribution frames of each class.

A more detailed analysis of our training data provides the first element of understanding. When we do not have sufficient data, it is impossible to carry out the initial application efficiently. In fact, the number of images for each class is not equivalent, so unbalanced classes are obtained, frequented classes are acquired, and rare classes are procured in case we have little data. This lack of data blocks our network in two different ways:

Table 6: Dataset anavid [21].

Method	Used dataset	
	Class	Number of frames
Proposed model 1	Angry	17,312
	Sad	18,542
	Neutral	21,140
	Happy	18,542
	Total	75,536
Proposed model 2	Angry	31,822
	Sad	30,128
	Neutral	37,179
	Happy	31,157
	Total	130,286
Proposed model 3	Angry	69,892
	Sad	68,053
	Neutral	69,124
	Happy	67,832
	Total	274,901

(i) By not providing it with sufficient information to allow it to learn to differentiate certain classes

(ii) By leading it to specialize in learning data (overfitting) and therefore be unable to provide a correct answer for the test data

For this reason, the neural network can face the overfitting problem. In brief, neural networks require a lot of learning data and balanced classes.

The model which is generated by training the abovementioned three datasets is tested on a test dataset that includes 2,000 images for each class (happy, neutral, angry, and sad). The result is illustrated in different confusion matrices (see Table 7).

It is interesting to look at class-by-class results through a confusion matrix. It indicates the proportion of data correctly predicted for each class through a diagonal column. The rest of the data is assigned to other classes (see Table 8). This shows a considerable improvement in the confusion matrix as it helps to generalize the model and boost its performance. Early stopping is utilized to avoid overfitting in our experiments. Throughout the simulation, the designed networks are trained with an iterative technique which allows the model to better fit the training dataset. Early stopping improves the model performance on any given data outside the training set.

4.7. Comparison with State-of-the-Art Methods. The proposed method is compared to previous works [11–17, 43]. Experiments demonstrate that our model outperforms the state-of-the-art methods on FER2013 and achieves high results on accuracy with 65.89% (see Table 8).

Table 7: Confusion matrices' comparison.

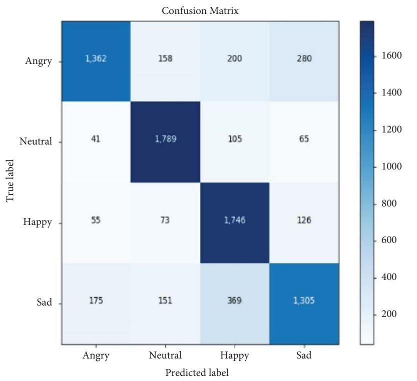
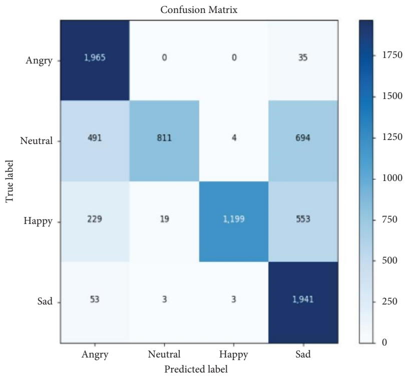
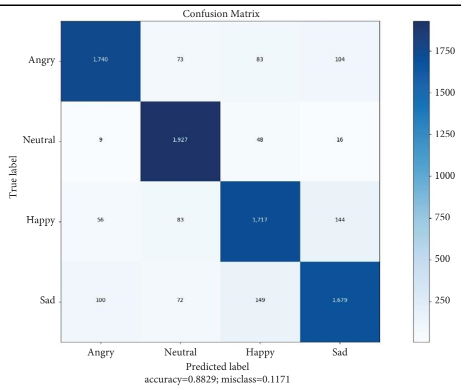
Method	Confusion matrix
Proposed model 1	
Proposed model 2	
Proposed model 3	

Table 8: Result on FER2013.

Method	Accuracy (%)	Gain (%)
AlMarri's method [15]	30.19	35.7
Mollahosseini et al.'s method [12]	61.10	4.79
Haque and Valles' method [11]	63.10	2.79
Zeng et al.'s method [43]	61.86	4.03
Xie et al.'s method [14]	62.99	2.90
Tümen et al.'s method [13]	57.10	8.79
Github et al.'s method [16, 17]	65.34	0.55
Proposed method	65.89	

Simulation results on FER2013 are listed in Table 8. They show that the proposed method provides greater results and a better performance gain from 0.55% to 35.7%.

5. Conclusion

Facial emotions are important factors in human communication. The feeling and the expression of emotions are the basic skills of social interaction. This work aims to study human emotion. FER can also be combined with early stopping and data augmentation using ordinary transformations and GAN technologies in order to boost the performance of our DL algorithm. Moreover, different experiments are applied to analyse the validity and the applicability of this technique with other methods in the field of emotion recognition. In future works, we will carry out new experiments to further ameliorate this research paper. In addition, we will further enhance the performance of the Jetson card by studying the structural characteristics of embedded systems and improving their functionality.

Data Availability

The data that support the findings of this study are available from the corresponding author upon reasonable request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

Acknowledgments

This work was performed under the MOBIDOC PromESsE fund from the Tunisian Ministry of Higher Education and Scientific Research. This research publication was supported by ANAVID France Company. The authors would like to express their sincere gratitude to ANAVID Company and LETI Research Laboratory.

References

[1] Y. Jayasimha and R. V. S. Reddy, “A facial expression recognition model using hybrid feature selection and support vector machines,” *International Journal of Information and Computer Security*, vol. 14, no. 1, pp. 79–97, 2021.

[2] I. Perikos, S. Kardakis, and I. Hatzilygeroudis, “Sentiment analysis using novel and interpretable architectures of Hidden Markov Models,” *Knowledge-Based Systems*, vol. 229, Article ID 107332, 2021.

[3] H. I. Dino and M. B. Abdulrazzaq, “Facial expression classification based on SVM, KNN and MLP classifiers,” in *Proceedings of the 2019 International Conference on Advanced Science and Engineering (ICOASE)*, pp. 70–75, IEEE, Zakho, Iraq, April 2019.

[4] A. Mostafa, M. I. Khalil, and H. Abbas, “Emotion recognition by facial features using recurrent neural networks,” in *Proceedings of the 2018 13th International Conference on Computer Engineering and Systems (ICCES)*, pp. 417–422, IEEE, Cairo, Egypt, December 2018.

[5] Y. Li, R. Wang, and Z. Yang, “Optimal scheduling of isolated microgrids using automated reinforcement learning-based multi-period forecasting,” *IEEE Transactions on Sustainable Energy*, vol. 13, no. 1, pp. 159–169, Jan. 2022.

[6] Z. B. Shi, T. Yu, Q. Zhao, Y. Li, and Y. B. Lan, “Comparison of algorithms for an electronic nose in identifying liquors,” *Journal of Bionics Engineering*, vol. 5, no. 3, pp. 253–257, 2008.

[7] D. A. Pitaloka, A. Wulandari, T. Basaruddin, and D. Y. Liliana, “Enhancing CNN with preprocessing stage in automatic emotion recognition,” *Procedia Computer Science*, vol. 116, pp. 523–529, 2017.

[8] I. Bendjoudi, F. Vanderhaegen, D. Hamad, and F. Dornaika, “Multi-label, multi-task CNN approach for context-based emotion recognition,” *Information Fusion*, vol. 76, pp. 422–428, 2021.

[9] A. Sinha and R. P. Aneesh, “Real time facial emotion recognition using deep learning,” *International Journal of Innovations and Implementations in Engineering*, vol. 1, 2019.

[10] R. Zatarain Cabada, H. Rodriguez Rangel, M. L. Barron Estrada, and H. M. Cardenas Lopez, “Hyperparameter optimization in CNN for learning-centered emotion recognition for intelligent tutoring systems,” *Soft Computing*, vol. 24, no. 10, pp. 7593–7602, 2020.

[11] M. I. U. Haque and D. Valles, “A facial expression recognition approach using DCNN for autistic children to identify emotions,” in *Proceedings of the 2018 IEEE 9th Annual Information Technology, Electronics and Mobile Communication Conference (IEMCON)*, pp. 546–551, Vancouver, Canada, December 2018.

[12] A. Mollahosseini, D. Chan, and M. H. Mahoor, “Going deeper in facial expression recognition using deep neural networks,” in *Proceedings of the IEEE Winter Conference on Applications of Computer Vision*, pp. 1–10, Waikola, HI, USA, March 2016.

[13] V. Tümen, Ö. F. Söylemez, and B. Ergen, “Facial emotion recognition on a dataset using convolutional neural network,” in *Proceedings of the 2017 International Artificial Intelligence and Data Processing Symposium (IDAP)*, pp. 1–5, IEEE, Malatya, Turkey, September 2017.

[14] S. Xie, H. Hu, and Y. Wu, “Deep multi-path convolutional neural network joint with salient region attention for facial expression recognition,” *Pattern Recognition*, vol. 92, pp. 177–191, 2019.

[15] S. B. S AlMarri, “Real-time facial emotion recognition using fast R-CNN,” Thesis, Rochester Institute of Technology, 2019.

[16] Github, “Real time human facial emotion recognition code,” 2022, https://github.com/shashikg/EmotionRecognitiondemo.

[17] S. Gupta and Shashi Kant, “Investigating emotion-color association in deep neural networks,” 2020, https://arxiv.org/abs/2011.11058.

[18] N. Polyzotis and M. Zaharia, “What can data-centric AI learn from data and ML engineering?” 2021, https://arxiv.org/abs/2112.06439.

[19] P. W. Wu, Y. J. Lin, C. H. Chang, E. Y. Chang, and S. W. Liao, “Relgan: multi-domain image-to-image translation via relative attributes,” in *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5914–5922, Montreal, Canada, October 2019.

[20] Github, “Github,” 2022, https://github.com/willylulu/RelGAN.

[21] Anavid, “Anavid,” 2022, https://www.anavid.co/.

[22] W. Xia, Y. Zhang, Y. Yang, J. H. Xue, B. Zhou, and M. H. Yang, “Gan inversion: a survey,” 2021, https://arxiv.org/abs/2101.05278.

[23] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” *IEEE transactions on affective computing*, vol. 13, 2020.

[24] E. Härkönen, A. Hertzmann, J. Lehtinen, and S. Paris, "Ganspace: discovering interpretable gan controls," 2020, https://arxiv.org/abs/2004.02546.

[25] J. Zhao, X. Mao, and L. Chen, "Speech emotion recognition using deep 1D & 2D CNN LSTM networks," *Biomedical Signal Processing and Control*, vol. 47, pp. 312–323, 2019.

[26] Y. Wei, F. Yang, and M. J. Wainwright, "Early stopping for kernel boosting algorithms: a general analysis with localized complexities," 2017, https://arxiv.org/abs/1707.01543.

[27] P. L. C. Courville, A. Goodfellow, I. J. M. Mirza, and Y. Bengio, *FER-2013 Face Database*, Universit de Montreal, Montréal, QC, Canada, 2013.

[28] E. Dufourq, "A survey on factors affecting facial expression recognition based on convolutional neural networks," in *Proceedings of the Conference of the South African Institute of Computer Scientists and Information Technologists*, pp. 168–179, Thaba Nchu, South Africa, September 2020.

[29] D. Lundqvist, A. Flykt, and A. Öhman, "The Karolinska directed emotional faces- kdef," in *CD ROM from Department of Clinical Neuroscience, Psychology Section*, pp. 1157–1161, Karolinska Institutet, Stockholm, Sweden, 1998.

[30] M. G. Calvo and D. Lundqvist, "Facial expressions of emotion (KDEF): identification under different display-duration conditions," *Behavior Research Methods*, vol. 40, no. 1, pp. 109–115, 2008.

[31] P. Lucey, J. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The extended cohn-kanade dataset (CK+): a complete dataset for action unit and emotion-specified expression," in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 94–101, New Orleans, LA, USA, June 2010.

[32] vision, "vision," 2023, http://vision.ucsd.edu/content/yale-face-database.

[33] The Mplab Genki Database, "GENKI-4K subset," 2023, http://mplab.ucsd.edu.

[34] M. Aza, Fasma'ul, N. Suciati, and S. C. Hidayati, "Performance study of facial expression recognition using convolutional neural network," in *Proceedings of the 6th International Conference on Science in Information Technology (ICSITech)*, IEEE, Palu, Indonesia, October 2020.

[35] I. Düntsch and G. Gediga, "Confusion matrices and rough set data analysis Journal of Physics: conference Series," *Journal of Physics: Conference Series*, vol. 1229, no. 1, Article ID 012055, 2019.

[36] developer, "developer," 2022, https://developer.nvidia.com/embedded/jetson-agx-xavier-developer-kit.

[37] michaelblogscode, "Michaelblogscode," 2022, https://michaelblogscode.wordpress.com/2017/10/10/reducing-and-profiling-gpu-memory-usage-in-keras-with-tensorflow-backend/.

[38] M. Li, X. Li, W. Sun, X. Wang, and S. Wang, "Efficient convolutional neural network with multi-kernel enhancement features for real-time facial expression recognition," *Journal of Real-Time Image Processing*, vol. 12, 2021.

[39] E. Correa, A. Jonker, M. Ozo, and R. Stolk, "Emotion recognition using deep convolutional neural networks," Technical Report IN4015, Cambridge University, Cambridge, CA, USA, 2016.

[40] T. Hu, W. Wang, C. Lin, and G. Cheng, "Regularization matters: a nonparametric perspective on overparametrized neural network," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, pp. 829–837, PMLR, Lauderdale, FL, USA, March 2021.

[41] G. S. Nandini, A. S. Kumar, A. P. S. Kumar, and K. Chidananda, "Dropout technique for image classification based on extreme learning machine," *Global Transitions Proceedings*, vol. 2, no. 1, pp. 111–116, 2021.

[42] S. H. Gao, Q. Han, D. Li, M. M. Cheng, and P. Peng, "Representative batch normalization with feature calibration," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 8669–8679, Salt Lake City, UT, USA, June 2021.

[43] G. Zeng, J. Zhou, X. Jia, W. Xie, and L. Shen, "Hand-crafted feature guided deep learning for facial expression recognition," in *Proceedings of the IEEE Conference on Automatic Face and Gesture Recognition (FG)*, pp. 423–430, Waikoloa Beach, HI, USA, May 2018.

[44] M. J. Lyons, "Excavating AI Re-excavated: debunking a fallacious account of the JAFFE dataset," 2021, https://arxiv.org/abs/2107.13998.

[45] Kaggle, "Mma facial expression," 2021, https://www.kaggle.com/mahmoudima/mma-facial-expression.