Decoding Emotions: How Temporal Modelling Enhances Recognition Accuracy

S. Chandrasekharan * **H. Arabian** ** **K. Moeller** **

* *Hochschule Furtwangen, VS-Schwenningen 78054, Germany.; e-mail: sch41871@stud.hs-furtwangen.de*
** *Institute of Technical Medicine (ITeM), VS-Schwenningen 78054, Germany.*

Abstract: Facial emotion recognition (FER) is a crucial component in the field of human-computer interaction, it can be strategically utilized for the therapeutic intervention of individuals who have been clinically diagnosed with autism spectrum disorder (ASD). The application of deep learning techniques for FER has the potential to enhance therapeutic intervention for this population. However, achieving high accuracy and rapid response times is essential for the practical implementation of emotion recognition systems. This paper introduces two models that incorporate temporal modelling, an approach that employs the time domain aspect, to enhance the accuracy of recognition. The results demonstrate that the integration of the time domain approach significantly improved the prediction accuracy, with an average increase of 8% compared to the convolution neural network (CNN) model. This research underscores the potential of long short-term memory (LSTM) in advancing the field of FER, leading to more efficient human-computer interaction systems.

Keywords: deep learning, emotion recognition, facial expression analysis, GoogLeNet, OULU-CASIA, temporal modeling.

1. INTRODUCTION

Facial expressions are a powerful medium for conveying emotional states. They provide key insights into an individual's current mood or mental state (Thiam et al., 2020). However, interpreting these expressions pose a significant challenge for individuals with autism spectrum disorder (ASD), who often struggle with task organization and sequencing. ASD is commonly diagnosed in children around the age of 3 years (Lord et al., 2000). A device capable of predicting changes in the emotional state of children with ASD could potentially aid caregivers and parents in preventing emotional outbursts (Bagirathan et al., 2021).

This study focuses on a time-based approach that considers sequences of emotions. Determining facial emotions from image sequences is a complex task due to various factors such as age, gender, frame background, image information scarcity, and clarity variations caused by changes in lighting conditions and poses (Singh et al., 2023). However, several successful approaches have been developed to determine facial emotion recognition (FER) from video sequences (Zhou et al., 2023; Singh et al., 2023; Thiam et al., 2020). To enhance recognition efficiency, various models incorporating recurrent neural networks (RNN) such as long short-term memory (LSTM) along with convolution neural network (CNN) have been introduced (Ming et al., 2022).

Recently, Zheng et al. conducted a study to enhance the recognition accuracy for finding a person's true feelings using facial micro-expressions.This was achieved by employing a combined neural network model. This model, developed by CNN and LSTM networks, was trained using a micro-expression dataset. The dataset included different emotions such as happiness, fear, anger, surprise, disgust, and sadness (Zheng and Blasch, 2023). In 2020, Mehendale conducted a study using a two-part CNN setup. This setup was used to get rid of background noise and for extraction of features. The model was trained using a database that had 10,000 images (Mehendale, 2020).

The research conducted by Singh et al. aimed to extract facial emotions from video clips. A three-dimensional convolutional neural network and LSTM were utilized for this purpose. The proposed hybrid architecture was designed to capture spatial information from video sequences of emotions. This model was tested using three publicly available FER datasets, namely SAVEE, CK+, and AFEW (Singh et al., 2023). Also, Arabian et al. explored the impact of network architecture on FER using various architectures such as VGG16, ResNet50, GoogleNet, ShuffleNet, and EfficientNetb0. The OULU-CASIA, FACES, and JAFFE datasets were used to establish a correlation between the accuracy and the features developed based on the architecture (Arabian et al., 2022).

In this study, the impact of a sequential approach on FER is analyzed. The CNN architecture of GoogleNet (Li et al.,

⋆ This research was partially funded by the German Federal Ministry of Research and Education (BMBF) under grant LESSON FKZ: 3FH5E10IA, a grant from KOMPASS funded by the Ministerium für Wissenschaft, Forschung und Kunst (MWK) of Baden-Wuerttemberg Germany, a grant from the ERAPERMED2022-276—ETAP BMG FKZ 2523FSB110, and a DAAD grant AIDE-ASD FKZ 57656657.

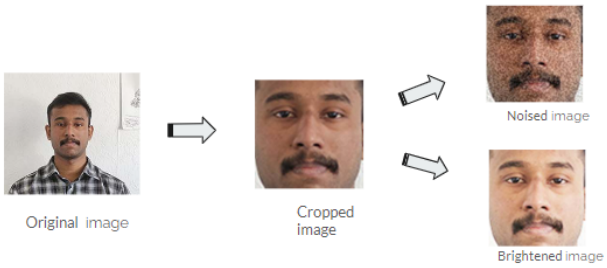

Fig. 1. Data pre-processing workflow of adding noise and brightness to images, original image (left), cropped image (middle), noised (right top) and brightened image (right bottom).

2021) trained on OULU-CASIA (Zhao et al., 2011) dataset serves as the base model. Two models were developed, one employed transfer learning and another incorporated LSTM, are studied and compared using the OULU-CASIA dataset.

2. METHODS

2.1 Methodology

This research formulates two models to assess the influence of LSTM layers on FER. The first model, referred to as Model 1, is developed using a transfer learning approach. It used pre-trained GoogleNet architectures for training. The second model, referred to as Model 2, is developed using the features extracted from Model 1 after its training. Model 2 incorporates LSTM layers and uses the features obtained from the last global average pooling layer of the network architecture used in Model 1 for its training. During the initial phase of Model 1, various image preprocessing techniques were applied.

The training process used various performance criteria. The OULU-CASIA database was used for this purpose, with the database split into 70% for training and 30% for validation.

2.2 Data pre-processing

The Viola-Jones (Viola and Jones, 2001) cascade detection algorithm was employed to filter the database and eliminate background artifacts. The local binary patterns method was utilized to select suitable faces for training, similar to the approach from (Arabian et al., 2021), resulting in 455 image sequences with varying length. The images are adjusted to dimensions of 224x224, depending on the input specifications of the GoogleNet architecture.

Subsequently, the images are injected with different types of noise prior to training and validation. The salt and pepper noise was added to the images to improve the performance of the model (Zhou et al., 2023). In addition to noise, the size of input database was increased by adding brightness of the original images by a factor of 1.5, as shown in Fig 1.

Table 1. Model 1 network training options

Parameter	Method / Value Selected
Solver	adaptive moment estimation (ADAM)
Mini Batch Size	50
# Epochs	40
Initial Learning Rate	0.0001

Table 2. Model 2 network training options

Parameter	Method / Value Selected
Solver	stochastic gradient descent with momentum (SGDM)
Mini Batch Size	50
# Epochs	40
Initial Learning Rate	0.0001

2.3 Network model architecture design

Two unique models, named Model 1 and Model 2, have been constructed to measure the efficiency of FER systems.

Model 1 The architecture of the model is based on GoogleNet, as depicted in Fig 2. GoogleNet is a deep learning CNN architecture that utilizes a 22-layer deep network, referred to as an inception module by its creators (Kalaiarasi and Esther Rani, 2021). This module is used to recognize complex patterns in the input data. In this model, the input data are images with facial expressions. The GoogleNet model processes these input images and classifies them into seven emotion class, and the output is taken from the *loss3-classifier* layer.

Model 2 The second model is based on LSTM. The LSTM networks are uniquely equipped to process data sequences with extensive dependencies. This is attributed to their distinctive architecture that incorporates gating mechanisms, such as the forget gate. These mechanisms enable LSTMs to regulate the flow of information and gradients throughout the network, thereby enhancing their performance in sequence processing tasks Vennerød et al. (2021). The model 2 uses the features of different emotion sequences extracted from the *pool5-7x7_s1* layer of GoogleNet from Model 1. Each emotion sequence has varying lengths. The features are input to the LSTM architecture to determine the effectiveness for video-based FER.

2.4 Performance Criteria

Model 1 and Model 2 are trained according to the parameters detailed in Tables 1 and 2, respectively. Model 1 uses pre-processed images as input, while Model 2 utilizes features from emotion sequences which is extracted from global average pooling layer. The dataset for both models were split into a 70% training and 30% validation set. The performance of the models is validated by evaluating the two highest true positive accuracies.

2.5 Database

In this study, the Oulu-CASIA database which consists of image frames was utilized for model training. The data is sourced from 80 individuals aged between 23 and 58,

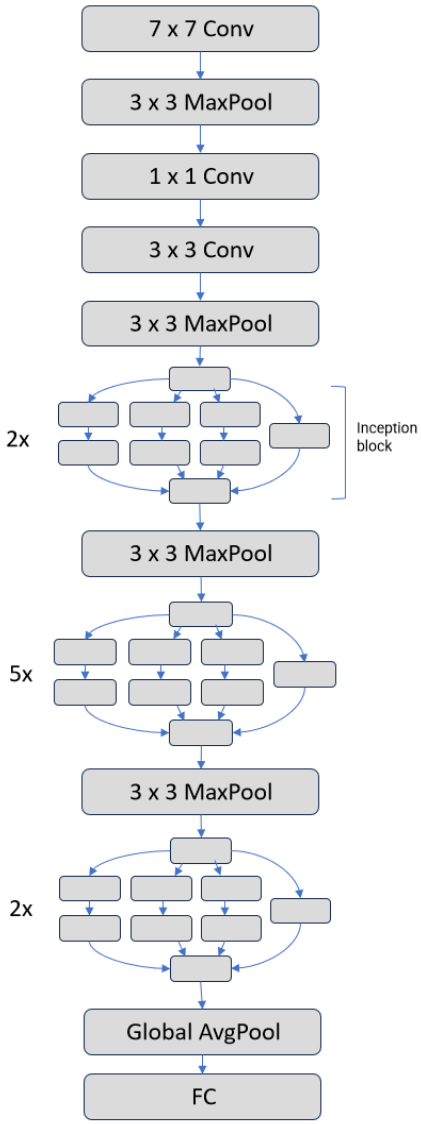

Fig. 2. GoogleNet architecture

each expressing six emotions such as anger, disgust, fear, happiness, sadness, and surprise. Emotional transitions are recorded from a neutral to an emotional state. Images under visible light and strong illumination were chosen, resulting in a total of 10,379 images (Zhao et al., 2011). The first three frames of each sequence were labeled as neutral, leading to seven emotion classes.

3. RESULTS & DISCUSSION

The performance of the models with GoogleNet was compared using three different sets of input images. Additionally, the impact of variation on recognition accuracy of Model 2 was analyzed with changing in number of output nodes.

3.1 Model performance

Three sets of input images were used for this study, the original image, a combination of the original and noised images, and a combination of the original, noised, and brightened images. Increasing the size of the dataset in this manner helped to mitigate the issue of overfitting during model training.

The performance of the models is illustrated in Fig. 3. For Model 1, the accuracy of predictions increased from 83.5% to 87.5% when the dataset was augmented with noisy images. This suggests that the model was able to generalize better with a more diverse set of training data. However, when the dataset was further augmented with brightened images, the accuracy increased from 83.5% to 84.5%. This was due to the model's inability to extract meaningful features from the brightened images.

On the other hand, Model 2, which incorporates the time domain and input noisy image, showed an increase in accuracy from 87.5% to 94.5% compared to Model 1. Similarly, the accuracy of the combined input dataset of original, noisy, and brightness-adjusted images increased from 84.5% to 93.5%. This substantial improvement was noticed because of the LSTM layer's ability to capture temporal dependencies in images sequence.

The accuracy of Model 2 was dependent on the precision of the features extracted from the Model 1. Model 1's predictions were more precise, it provided a more robust basis for Model 2, which in turn enhanced its performance. Conversely, less performance in Model 1's predictions propagated to Model 2, affected its ability to make correct classifications. Furthermore, it was observed a reduction in accuracy when additional variations, such as brightness adjustments, were introduced to the dataset. This decline can be attributed to more complexity resulting from the combining of diverse variations, which presents challenges for the model in discerning features effectively.

An ablation study was conducted to evaluate the impact of varying the number of LSTM nodes on the models' performance. The results, as depicted in Fig. 4, indicate that the performance of the models remained relatively stable across different LSTM configurations. Specifically, the models with LSTM nodes (64, 128, 250) demonstrated similar levels of high accuracy, suggesting a plateau effect beyond a certain number of nodes. This observation implies that the current configuration of the LSTM layer may already be near-optimal for this specific task. However, it's worth noting an anomaly in the data point for LSTM with 128 nodes under 'noise', which may warrant further investigation.

Given the observed improvement with the use of LSTM layers, future work could focus on exploring this approach with different datasets and varying the architecture of the layers. This could potentially lead to further enhancements in the performance of FER systems.

Moreover, live testing is required for further validation of the models. This will ensure that the models perform well in real-world scenarios and not just in controlled environments. The live testing could provide valuable insights into the practical applicability and effectiveness of the models, thereby guiding future improvements.

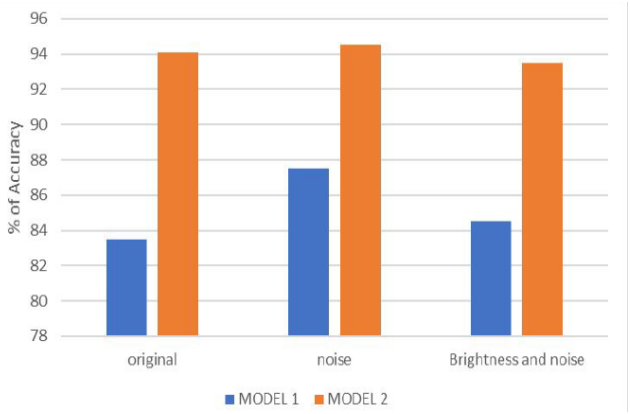

Fig. 3. Comparison of Model 1 and Model 2 with GoogleNet architecture using the different image inputs.

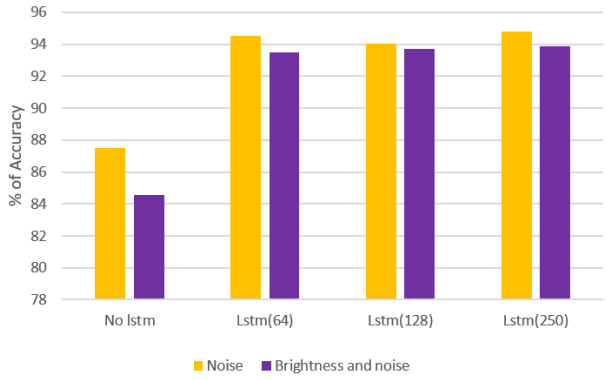

Fig. 4. Performance Accuracy Comparison with different number of LSTM nodes using two image input methods.

4. CONCLUSION

In this study, the impact of the temporal aspect on the performance of a CNN model for FER was analyzed. The results demonstrated that incorporating a sequential approach using long short-term memory into a CNN model improved performance by an average of 8%. It was also found that the recognition accuracy of the CNN, was not significantly impacted by the number of nodes in the LSTM layer.

REFERENCES

Arabian, H., Wagner-Hartl, V., Chase, J.G., and Möller, K. (2021). Image pre-processing significance on regions of impact in a trained network for facial emotion recognition. *IFAC-PapersOnLine*, 54(15), 299–303.

Arabian, H., Wagner-Hartl, V., and Moeller, K. (2022). Network architecture influence on facial emotion recognition. In *Current Directions in Biomedical Engineering*, volume 8, 524–527. De Gruyter.

Bagirathan, A., Selvaraj, J., Gurusamy, A., and Das, H. (2021). Recognition of positive and negative valence states in children with autism spectrum disorder (asd) using discrete wavelet transform (dwt) analysis of electrocardiogram signals (ecg). *Journal of Ambient Intelligence and Humanized Computing*, 12, 405–416.

Kalaiarasi, P. and Esther Rani, P. (2021). A comparative analysis of alexnet and googlenet with a simple dcnn for face recognition. In *Advances in Smart System Technologies: Select Proceedings of ICFSST 2019*, 655–668. Springer.

Li, S., Li, W., Wen, S., Shi, K., Yang, Y., Zhou, P., and Huang, T. (2021). Auto-fernet: A facial expression recognition network with architecture search. *IEEE Transactions on Network Science and Engineering*, 8(3), 2213–2222. doi:10.1109/TNSE.2021.3083739.

Lord, C., Cook, E.H., Leventhal, B.L., and Amaral, D.G. (2000). Autism spectrum disorders. *Neuron*, 28(2), 355–363.

Mehendale, N. (2020). Facial emotion recognition using convolutional neural networks (ferc). *SN Applied Sciences*, 2(3), 446.

Ming, Y., Qian, H., Guangyuan, L., et al. (2022). Cnn-lstm facial expression recognition method fused with two-layer attention mechanism. *Computational Intelligence and Neuroscience*, 2022.

Singh, R., Saurav, S., Kumar, T., Saini, R., Vohra, A., and Singh, S. (2023). Facial expression recognition in videos using hybrid cnn & convlstm. *International Journal of Information Technology*, 15(4), 1819–1830.

Thiam, P., Kestler, H.A., and Schwenker, F. (2020). Two-stream attention network for pain recognition from video sequences. *Sensors*, 20(3), 839.

Vennerød, C.B., Kjærran, A., and Bugge, E.S. (2021). Long short-term memory rnn. *arXiv preprint arXiv:2105.06756*.

Viola, P. and Jones, M. (2001). Rapid object detection using a boosted cascade of simple features. In *Proceedings of the 2001 IEEE computer society conference on computer vision and pattern recognition. CVPR 2001*, volume 1, I–I. Ieee.

Zhao, G., Huang, X., Taini, M., Li, S.Z., and PietikäInen, M. (2011). Facial expression recognition from near-infrared videos. *Image and vision computing*, 29(9), 607–619.

Zheng, Y. and Blasch, E. (2023). Facial micro-expression recognition enhanced by score fusion and a hybrid model from convolutional lstm and vision transformer. *Sensors*, 23(12), 5650.

Zhou, D., Cheng, Y., Wen, L., Luo, H., and Liu, Y. (2023). Drivers' comprehensive emotion recognition based on ham. *Sensors*, 23(19), 8293.