EasyChair Preprint

№ 8395

Real-Time Facial Emotion Recognition for Visualization Systems

Ceren Ozkara and Pïnar Oğuz Ekim

EasyChair preprints are intended for rapid dissemination of research results and are integrated with the rest of EasyChair.

July 5, 2022

Real-Time Facial Emotion Recognition for Visualization Systems

Ceren Özkara
Izmir University of Economics
Izmir, TURKEY
ceren.ozkara@std.izmirekonomi.edu.tr

Pınar Oğuz Ekim
Izmir University of Economics
Izmir, TURKEY
pinar.ekim@ieu.edu.tr

***Abstract*— This project aims to review the most popular deep learning algorithms and their performances in camera systems based on real-time facial emotion recognition and suggest a new model for future applications. Firstly, convolutional neural network (CNN) algorithms that recognize human emotions, such as AlexNet, GoogleNet, and VGG19, are investigated according to their performances. Then, the CNN algorithm with the best numerical performance is chosen for enhancement. After, the new hybrid model is constructed via chosen CNN and long short-term memory (LSTM). Lastly, the proposed model and face images achieved from the camera are combined to simulate real-time application.**

Keywords — face detection, facial expression recognition, CNN, LSTM, hybrid model

I. INTRODUCTION

People commonly use facial expressions to show their emotional states and make a communication. Researchers that know the importance of this nonverbal information for clear communication have started to develop emotion recognition systems based on facial expressions. These systems developed with deep learning algorithms and feature extraction methods have many usage and application areas, especially security, health, and human-machine interfaces [1].

In the recent past, investigations have focused on reconstructing the deep learning algorithms or combinations [2] [3]. Some researchers have proposed a model based on single deep convolutional neural networks for facial emotion recognition. The proposed model consists of 6 convolutional layers, three max-pooling layers after the convolutions layer, two deep residual learning boxes implemented after the second and fourth convolution layer, and two fully connected layers. The model performance has been tested on two public datasets: Extended Cohn–Kanade (CK+) and Japanese Female Facial Expression (JAFFE) [2]. The searchers interested in the health sector focus on the importance of automatic health issue detection because of the increasing quick result requirement in the vast population. Because of this reason, they suggest the CNN-LSTM hybrid model instead of the CNN approach. Obtained better validation accuracy and fast training result in the hybrid model shows that researchers will be focusing on the hybrid models in the future [3]. Researchers have analyzed four CNN architectures (GoogleNet, ResNet, VGGNet, and AlexNet) for facial expression recognition according to their validation accuracy on FER2013, which is one of the famous datasets. The original method for facial expression recognition is reconstructed to obtain basic structure, and this new model is shown to improve the accuracy result [4].

Moreover, investigated literature shows that there are so many datasets that are related to facial emotion detection. In the human-robot interactive processes, the main point is communication; therefore, researchers (Jaiswal and Nandi) focused on the applications that allow the robots to understand the human facial expression. They offered a CNN-based model and tested it over the eight various datasets: Fer2013, CK, CK+, Chicago Face Database, JAFFE Dataset, FEI face dataset, IMFDB, and TFEID [5].

Researchers (Gao et al.) who know feature extraction is important in deep learning used VGG-19 architecture in camera systems. Additionally, they showed that systems that applied transfer learning could extract more information than the original network while reducing the training time [6].

II. METHODOLOGY

The methodology section explains the deep learning algorithms and their usage in the project. Moreover, this section clarifies the requirements for implementation.

A. CNN

A Convolutional Neural Network (CNN, or ConvNet) are multi-layer neural network designed to identify visual structures directly from pixel images with reprocessing. The layers of the network and their purposes are as follows [7].

- *Convolutional Layer:* The convolutional layer is used for extracting high-level and low-level features by using the filter.
- *Non-Linearity Layer:* The non-linearity layer is used for introducing the non-linearity to the system. It is also called the activation layer. Sigmoid, tanh, and ReLU are the most popular functions in this layer.
- *Pooling Layer:* The pooling layer is used for downsampling network parameters and reducing the calculation. Also, the pooling layer checks the suitability of the network. Max pooling is the most popular algorithm; moreover, average pooling and L2-norm pooling are also used.
- *Flattening Layer:* The flattening layer is used for preparing the input of the classical neural network. The input is converted to a vector.
- *Fully Connected Layer:* The fully connected layer is used for learning via a neural network.

This article uses the AlexNet, GoogleNet, and VGG architectures are prevalent CNN models.

1) AlexNet

In 2012, Alex Krizhevsky, Ilya Sutskever, and Geoffrey Hinton developed AlexNet architecture based on CNN, which is described in figure 1. AlexNet consists of 5 convolutional layers and three fully connected layers. The input of the network is 227by 227-pixel RGB images. Although standard neural networks use tanh or sigmoid,

AlexNet uses ReLU (Rectified Linear Unit) as activation in non-linear parts. ReLU function has a time advantage when the model is being trained. Also, sigmoid function derivative approaches zero and causes a vanishing gradient. Therefore, it becomes difficult to update the weights in the model. AlexNet uses max-pooling in pooling layers to reduce the number of calculated parameters in the network. Approximately 60 million parameters are calculated with AlexNet. AlexNet completed the ImageNet Large Scale Visual Recognition Challenge with a 15.3% error rate [8].

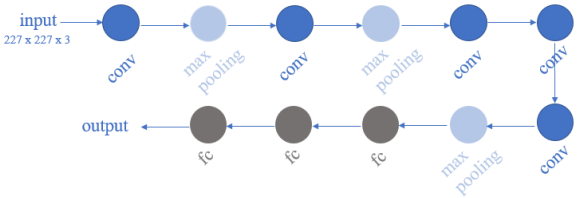

Figure 1: *AlexNet Architecture*

2) GoogleNet

GoogleNet, developed by researchers at Google, consists of 22 layers (27 layers including pooling layers), as in figure 2. The nine inception modules provide feature detection through convolutions with different filters at different scales. Moreover, inception module illustrated in figure 3 reduces the computational cost of training an extensive network through dimensional reduction. The average pooling layer takes a mean from all the feature maps produced by the last inception module. A dropout layer prevents overfitting the network and is used just before the linear layer. The dropout technique randomly reduces the number of interconnecting neurons within a neural network. The connected neurons are reduced randomly in this layer. GoogleNet architecture's last two layers are the linear layer which consists of 1000 hidden units, and the softmax layer, which uses the softmax function. The GoogleNet architecture was performed in the ILSVRC 2014 classification challenge with an error of 6.67% [9-10].

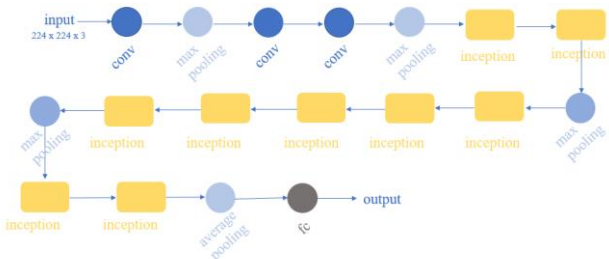

Figure 2: *GoogleNet Architecture*

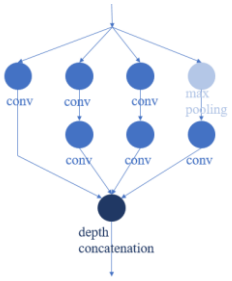

Figure 3*: Inception Module*

3) VGG-19

Visual Geometry Group develops VGG-19 at Oxford. VGG-19 architecture consists of a total of 24 main layers, which are 16 convolutional, five pooling, and three fully connected layers, as in figure 4. Furthermore, it contains approximately 138 million parameters. The network has an image input size of 224-by-224 RGB [11].

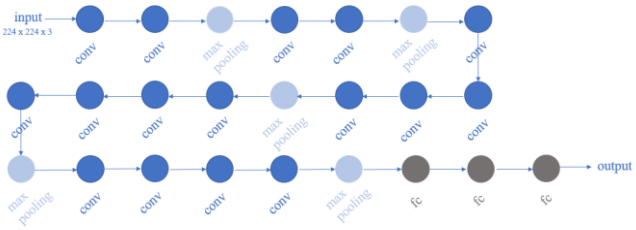

Figure 4: *VGG-19 Architecture*

B. Transfer Learning

The increasing dataset and complex model architecture cause difficulties in using the standard computer while the training process. Especially training process that takes days or sometimes weeks makes it impossible to work on standard computer processors. Learning from scratch, which is commonly performed for each learning process, causes vital problems in these years because of spending time in training. As a result, the method that is called transfer learning was developed. In this way, learned information from some tasks in other tasks will be possible and advantageous to use in other tasks. In other words, information such as features, and weights, obtained from previously trained models can be used for a new task [12]. The working principle of transfer learning is summarized in the following figure 5 to provide a better understanding.

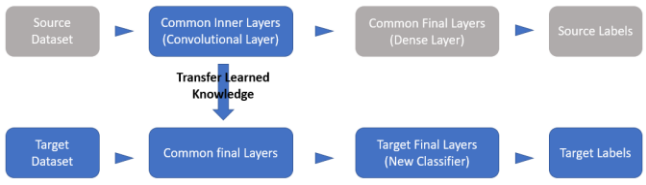

Figure 5: *Transfer Learning*

C. LSTM

LSTM is a special type of recurrent neural network (RNN) architecture developed by Hochreiter and Schmidhuber [13]. The LSTM architecture consists of sequential blocks that repeat each other. These sequential blocks are input gate, a “forget” gate, and an output gate. Firstly, information that will be forgotten is specified by using the input. These are done in the forget gate. Secondly, information is updated in the input gate by using sigmoid as the activation function. Then, the new information is generated with tanh function. The process described continues iteratively to minimize the difference between the actual training values and the LSTM output values [14].

D. Hybrid Model

The CNN architecture is not suitable for continuous dynamic images, and RNN has an internal memory to process dynamic data. Thus, CNN and LSTM combinations can be used to obtain a more beneficial model. In the offered hybrid model, CNN is used for deep feature extraction, and LSTM is used for a classifier. [20-21-22].

E. System Model

The model of the system is illustrated in figure 6. The image is obtained from the camera. The face in the image is caught by the detection algorithm. The catch image is cropped and resized. Then, this is given as an input to the learning algorithm that is trained to recognize seven human emotions such as anger, disgust, fear, happiness, neutral, sad, and surprise. Lastly, classification output is taken as a facial expression result.

Figure 6: *System Model*

Viola-Jones algorithm, which is a popular detection model, is preferred as a detector in the system. It uses rectangular features to identify the particular object in the image. In this way, the system trained for face recognition is not interested in non-face areas [15].

III. Implementation Details

The implementation details, such as the feature of the dataset and used program, are given in section 3.

A. Dataset

FER2013[16] is the most popular dataset for facial emotion recognition implementation because of the number of images in the table 1. FER 2013 consists of 35887 grey images which are 48 by 48-pixel. The train set and test set have 28709 and 7178 images, respectively. There are seven emotions: anger, disgust, fear, happiness, neutral, sad, and surprise.

	Angry	***Disgust***	***Fear***	***Happy***	***Neutral***	***Sad***	***Surprise***
Train	3995	436	4097	7215	4965	4830	3171
Test	958	111	1024	1774	1233	1247	831

Table 1: *Number of Images in FER2013*

B. Simulation Environment

Analyses and real-time implementation were done in MATLAB2021a. The related deep learning toolboxes were initialized before the usage. Because of the high training speed of the graphics processing unit (GPU) [6], neural networks were trained using the NVIDIA GeForce840M.

IV. Results and Discussions

The performance results for architectures AlexNet, GoogleNet, VGG-19, and CNN-LSTM hybrid model, the adequacy of the dataset for the implementation, and the applicability of the real-time systems are examined in the results and discussions section.

The validation accuracy used for performance criteria explains the generalization or classification ability of the model. The obtained validation accuracy results for CNN architectures (GoogleNet, AlexNet, and VGG-19) that were trained via transfer learning are in table 2. The VGG-19 has the highest result with 62.66%, while the GoogleNet and AlexNet have 60.7% and 57.67%. As a result, the VGG-19 is the best option for implementation.

The elapsed time in training is another vital performance parameter. Table 2 shows that GoogleNet, AlexNet, and

Algorithm	***Validation Accuracy***	***Elapsed Time***	***Epoch Number***
GoogleNet	60.70 %	652 min 44 sec	6
AlexNet	57.67 %	353 min 6 sec	6
VGG-19	62.66 %	3697 min	6

Table 2: *Validation accuracy for CNN algorithms*

VGG-19 spent nearly 11 hours, 6 hours, and 2.5 days in the training, respectively. Therefore, the AlexNet is the most suitable model for applications.

GoogleNet validation accuracy result is not less than AlexNet, and elapsed time in the training is not too much like VGG-19. Thus, the GoogleNet algorithm was chosen to convert the stationary to a dynamic model. The CNN-LSTM hybrid model provided this improvement. The obtained performance result for the CNN-LSTM model is in table 3. The spending time in the training is increased by 11% compared with the original GoogleNet. Additionally, the reconstructed model showed a less than 1% decrease in the validation accuracy. As a result, the proposed hybrid model is transformed into the appropriate approach for real-time simulation.

Algorithm	***Validation Accuracy***	***Elapsed Time***	***Epoch Number***
CNN-LSTM	60.34 %	724 min 55 sec	6

Table 3: *Validation accuracy for hybrid model*

The suitability of the FER2013 is investigated using test images inside the dataset. When the performance of the trained algorithm is examined on the test dataset, the results for each emotion individually are in the following figure 7.

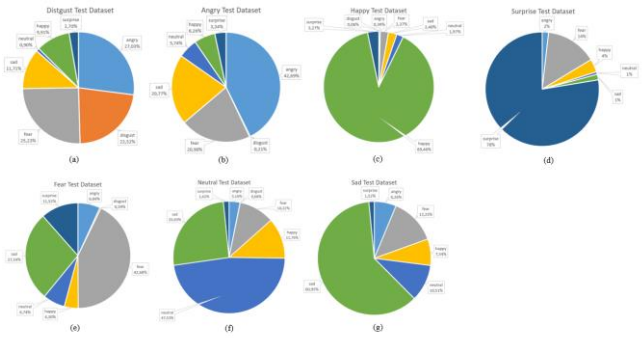

Figure 7:*Test Dataset Performance Results for (a) Disgust(b) Angry (c) Happy (d) Surprise (e) Fear (f) Neutral (g) Sad*

In figure 7(a), the test dataset results for the emotion of disgust are 27% for anger, 25% for fear, and %22 for disgust. The model was trained with fewer samples for disgust than other emotions, as seen in table 1. Thus, the system does not catch the disgust effectively [17]. If the emotion of anger is

examined for test dataset results, the highest results are 42% for anger, and 20% for both fear and sadness, as seen in figure 7(b). In figure 7(c), 89% of the test dataset is labeled happy, and the other emotions are less than 4%. When the emotion of surprise is analyzed in figure 7(d), the highest two results of the test dataset are 78% for surprise and 14% for fear. Additionally, the test dataset results of the fear show the highest results as 42% for fear, 27% for sad, and 11% for a surprise in figure 7(e). The similarity in facial features confuses the analysis of different emotions and causes errors. As a result, fear analyses have a surprise, and surprise analyses have fear [18]. In figure 7(f), the emotion of neutral has 47% of neutral results and 25% of sad results when the test dataset is researched. Figure 7(g) shows 60% sad and 13% fear results when the test dataset of sad emotions is examined. In conclusion, the model's validation and test accuracy are directly affected by the datasets with unbalanced and unclear labeled samples. Although FER2013 is considered a huge dataset, it is not a well-defined dataset because of these challenges [19].

Figure 8: *Simulation Results*

The results are as shown in figure 8 when the whole system is assembled for simulation as in the flow chart. Though FER2013 is not suitable for training and validation accuracy is 60.34 %, the proposed real-time system can give the wanted outputs.

V. Conclusion

In this paper, CNN architectures, which are AlexNet, GoogleNet, and VGG-19, were investigated according to validation accuracy performance for emotion recognition. The GoogleNet, which has the average validation accuracy and spending time, was used to construct the CNN-LSTM hybrid model. The constructed model showed similar performance results to analyzed CNN algorithms according to validation accuracy. The recommended model was used to implement real-time simulation. The hybrid models are the way that achieves various aims such as better performance and converting the domain from static to dynamic when the construction of the algorithms can allow combining parts of the different algorithms. In conclusion, the hybrid model approach can be used as a resource for real-time automated camera systems.

References

[1] X. Lu, “Deep Learning based emotion recognition and visualization of Figural Representation,” *Frontiers in Psychology*, vol. 12, 2022.

[2] D. K. Jain, P. Shamsolmoali, and P. Sehdev, “Extended Deep Neural Network for facial emotion recognition,” *Pattern Recognition Letters*, vol. 120, pp. 69–74, 2019.

[3] M. Z. Islam, M. M. Islam, and A. Asraf, “A combined deep CNN-LSTM network for the detection of novel coronavirus (COVID-19) using X-ray images,” *Informatics in Medicine Unlocked*, vol. 20, p. 100412, 2020.

[4] Y. Gan, “Facial expression recognition using convolutional neural network,” Proceedings of the 2nd International Conference on Vision, Image and Signal Processing, 2018.

[5] S. Jaiswal and G. C. Nandi, “Robust real-time emotion detection system using CNN architecture,” *Neural Computing and Applications*, vol. 32, no. 15, pp. 11253–11262, 2019.

[6] Z. Gao, Y. Zhang, and Y. Li, “Extracting features from infrared images using convolutional neural networks and transfer learning,” *Infrared Physics & Technology*, vol. 105, p. 103237, 2020.

[7] T. Ergin, “Convolutional Neural Network (convnet yada CNN) Nedir, Nasıl çalışır?,” Medium, 22-Feb-2020. [Online]. Available: https://medium.com/@tuncerergin/convolutional-neural-network-convnet-yada-cnn-nedir-nasil-calisir-97a0f5d34cad. [Accessed: 23-Jun-2022].

[8] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks." *Advances in Neural Information Processing Systems 25 (NIPS 2012)*, 2012.

[9] R. Alake, “Deep learning: Googlenet explained,” *Medium*, 03-Nov-2021. [Online]. Available: https://towardsdatascience.com/deep-learning-googlenet-explained-de8861c82765. [Accessed: 17-Jun-2022].

[10] C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[11] K. Simonyan, and A. Zisserman, "Very deep convolutional networks for large-scale image recognition." arXiv preprint arXiv:1409.1556 (2014).

[12] Ayyüce Kızrak, “Derine Daha DERİNE: Evrişimli Sinir Ağları,” *Medium*, 07-Jan-2020. [Online]. Available: https://ayyucekizrak.medium.com/deri%CC%87ne-daha-deri%CC%87ne-evri%C5%9Fimli-sinir-a%C4%9Flar%C4%B1-2813a2c8b2a9. [Accessed: 17-Jun-2022].

[13] Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. Neural computation, 9(8), 1735- 1780

[14] A. KARA, “Global Solar Irradiance Time Series Prediction Using Long Short-Term Memory Network.” [Online]. Available: https://dergipark.org.tr/en/download/article-file/878498. [Accessed: 17-Jun-2022].

[15] P. Viola and M. Jones, “Rapid object detection using a boosted cascade of Simple features,” Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. CVPR 2001.

[16] https://www.kaggle.com/msambare/fer2013

[17] H.-D. Nguyen, S. Yeom, G.-S. Lee, H.-J. Yang, I.-S. Na, and S.-H. Kim, “Facial emotion recognition using an ensemble of multi-level convolutional Neural Networks,” *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 33, no. 11, p. 1940015, 2019.

[18] E. Dandıl and R. Özdemir, “Real-time facial emotion classification using deep learning,” *Data Science and Applications*, vol. 2,no.1, pp. 13–17, 2019.

[19] P. Naga, S. D. Marri, and R. Borreo, “Facial emotion recognition methods, datasets and technologies: A literature survey,” *Materials Today: Proceedings*, 2021.

[20] A. Khamparia, B. Pandey, S. Tiwari, D. Gupta, A. Khanna, and J. J. Rodrigues, “An integrated hybrid CNN–RNN model for visual description and generation of captions,” *Circuits, Systems, and Signal Processing*, vol. 39, no. 2, pp. 776–788, 2019.

[21] I. E. Livieris, E. Pintelas, and P. Pintelas, “A CNN–LSTM model for gold price time-series forecasting,” *Neural Computing and Applications*, vol. 32, no. 23, pp. 17351–17360, 2020.

[22] T.-H. S. Li, P.-H. Kuo, T.-N. Tsai, and P.-C. Luan, “CNN and LSTM based facial expression analysis model for a humanoid robot,” *IEEE Access*, vol. 7, pp. 93998–94011, 2019.