VGG-16 and Optimized CNN for Emotion Classification

Amornvit Vatcharaphrueksadee* Rattikarn Viboonpanich*
Puttakul Sakul-ang** and Maleerat Maliyaem**

Received : November 19, 2020
Revised : December 16, 2020
Accepted : December 23, 2020

Abstract

This paper discusses the efficiency of using VGG 16 and Optimized CNN on emotion classification of human face dataset. Facial expression can be used as a communication medium between people to express a person feeling not only what it has shown on the outside also include the inner feeling, mental situation and perspective. In this paper, 5 basic emotions have been chosen to test the highly efficient models to elaborate the model complexity of the state-of-the-art model towards its accuracy and training efficiency of the proposed models against the state-of-the-art model such as VGG16 to achieve at least 65% accuracy.

Keywords: Emotion Recognition, Convolution Neural Network, Facial Action Coding System, Facial Action Unit.

1. Introduction

Understanding human emotions is a difficult task for the machine. Facial Action Coding System (FACS) is an anatomical analysis of human face movement developed by Ekman and Frieson [1] and further improve by Hager [2] to detect all possible identifiable movement of the facial features towards human emotions. The FACS can help the machine to understand human emotion based on facial movement analysis. Building machine that understands human has open opportunities for mainly human-computer interaction [3] e.g. better communication and interaction between human and machine. The current techniques to identify human emotions include using visual information such as facial features detection, body postures, audio information e.g. speed and ton of the voice, and medical information such as Electrocardiogram (EKG). In this paper, the focus information is the visual recognition of emotion due to solid information on classifying distinct emotions, which can be applied to human across the world [4]. Ekman and Friesen have classified human basic emotions into 6 basic emotions including anger, disgust, fear, happiness, sadness and surprise. This information, then, has been used by computer vision researchers in the task of identifying facial expression towards emotions or true feeling [5]. This research chooses only four of six basic emotions to test the state-of-the-art models against the proposed model including anger, happiness, sadness and surprise with the addition of neutral emotion. The emotion classification for the machine has potential in a variety of fields related to interaction and customer experiences. To achieve set goals, VGG-16 classifier baseline model and optimized convolution neural network (op-CNN) have been introduced to classify 5 basic emotions. The op-CNN is an optimized convolution neural network, which includes various deep learning techniques [6]. A couple of datasets has been used to leverage this research including CK+48 and FER2013 [7, 8]. To evaluate the performance of each model, training accuracy and validation accuracy have been measured. The op-CNN is expected to achieve a higher score compared to the traditional CNN with at least 65% of model accuracy.

2. Theoretical Background and Related Researches

2.1 Emotion

Emotion Recognition is a research area widely investigated in the Machine Learning field. Many researchers try to find solutions for correctly predicting human basic

*Faculty of Information Technology and Digital Innovation, North Bangkok University.
** Faculty of Information Technology and Digital Innovation, King's Mongkut University of Technology, North Bangkok.

emotions which can be useful for many innovative areas such as robotic, human-computer interaction (HCI), etc.

Human basic emotion has been studied and categorized into discrete categories using the language from daily life. The most popular examples are the 6 basic emotions proposed which are happiness, sadness, surprise, fear, anger, and disgust. Furthermore, it is also suggested that the facial expressions of these basic emotions are perceived in the same manner for all mankind.

Besides facial expression, human emotional states can be detected from other non-verbal sources and verbal sources. Body gestures also contain lots of information about the movement that can be categorized into different emotions. Wording and tone of voice are also as crucial as unspoken sources for classifying emotion as well.

2.2 Emotion Recognition

Tarnowski et al. [9] had proposed a method to detect human basic emotions with neutral. Microsoft Kinect, 3D face modelling, Action Units and neural networks were used in this research to classify emotions from input sources.

Microsoft Kinect is a low budget device for 3D face modelling. Despite having a low scanning resolution, this device can capture image sequences at 30 frames/seconds. Moreover, it is also equipped with an infrared emitter and double cameras for recording light and depth measurement. The Microsoft Kinect uses infrared emitters to capture reflected rays of capturing the subject to build a 3D model based on 121 specific points [10]. These points contain 3D camera-ray distance data (x, y, z) which are the distance between the camera and capturing the subject in specific areas on the face, e.g. nose, mouth, eyebrows, eyes, etc. Movement of a set of points is used for calculating special coefficients, called "Action Units (AU)".

In Microsoft Kinect, there are 6 Action Units (AU) extracted from the Facial Action Coding System (FACS). Each Action Unit represents movements of upper lip raising (AU0), jaw lowering (AU1), lip stretching (AU2), lowering eyebrows (AU3), lip corner depressing (AU4), and outer brow-raising (AU5). These movement data can be useful for describing the emotions of the human face.

To collect experimental data, in-lab experiments is a good procedure done by asking participants to mimic a face shown on the screen for 5 seconds. The on-screen faces were sampled from KDEF database [11], which are labelled into 7 basic emotions including neutral. Each participant was asked to perform 2 sessions. As a result, there are 252 facial expressions collected from the experiment arranged in the hierarchical structure, which will be fed into the system for classifying emotions.

Action Units (AU) can be used to map emotional states based on coefficient ranging between -1 to 1 labelled into neutral, joy, surprise, anger, sadness, fear, and disgust. Since, the spatial distribution of Action Units (AU) related to emotional states, K-Nearest-Neighbor (KNN) with any multi-dimensional machine learning classifier can be applied for automatically emotion recognition. However, there are ambiguities in the dataset, which are participant's response time and facial expression preparation occurred only in the first second of the dataset then starting to fade into the neutral so the clipping method has been applied to the train dataset for improve model efficiency.

Computer graphic technique of using 3D modeling has been introduced to emotion recognition by generating 3D faces regarding emotional states. This technique can improve the classifier results since 3D model can be rendered in any angle.

Disgust and fear are the most difficult emotion to distinguish based on the emotion recognition accuracy on 4 visual classifiers including Semi-supervised learning with 3D Auto encoder (S3DAE), Convolutional 3D with Auxiliary Network (C3DA), Parallel CNN and Landmark with the highest score of 5.0% and 4.35% as shown in table 1 [12].

In sum, building an emotion recognition model with good accuracy requires a careful selection of dataset, feature dropout, feature selection, and classifier.

Table 1. *Per-class emotion recognition accuracy based on visual classifiers.*

Network	Ang.	Dis.	Fear	Hap	Neu	Sad	Sur.
S3DAE	51.56	2.50	2.17	17.46	58.73	32.79	4.35
C3DA	46.88	0.00	4.35	49.20	65.08	9.83	34.78
Parallel CNN	45.31	5.00	4.35	57.14	73.02	19.67	19.56
Landmark	57.37	0.00	0.00	35.71	17.42	28.26	3.33

2.3 Literature Review

Using machine learning to predict emotion has been studied over recent years on both modal and multi-modal systems, which the multi-modal has proven to be more effective but more time-consuming in the training procedure. The datasets used to predict emotion from previous works on machine learning and neural networks to predict emotion are ranging from audio-visual expression to body gestures. Speech-based emotion classification [13] using Mel-frequency speech power coefficients and Hidden Markov Model has achieved higher accuracy of 94.44% and 70% on six basic emotion classification task. EEG was implemented with common spatial patterns (CSP) and linear-SVM to classify emotion with 93.5% accuracy using 1 to 3 seconds of EEG signal [14]. Audio-video emotion classification in EmotiW2017 challenge has achieved the classification accuracy of 60.03% with an optimized VGG16 [15]. 2D and 3D Convolutional Neural Network has been implemented on video emotion classification in the Wild with 58.8% of accuracy [16]. Mouth feature extraction with random trees, k-nearest neighbors, and multilayer-perceptron and support vector machine has been implemented towards emotion classification with 70% of accuracy over a single emotion [17]. The Multi-modalities Emotion recognition was investigated using Bayes Net and has achieved a better recognition accuracy than single modality [18].

3. Research Methodology

3.1 Convolutional Neural Network

Convolution Neural Network is a popular method for image classification that has been chosen with data augmentation in this research. Dataset used are FER2013 dataset and CK+48 dataset. The FER2013 dataset consists of 35,887 48x48 pixel grayscale images, which have been labelled with 7 emotions including Angry, Disgust, Fear, Happy, Sad, Surprise and Neutral. The CK+48 is a dataset with 5 emotions including anger, fear, happiness, sadness and surprise. As mentioned in the previous section, the disgust and fear are emotional states causing ambiguous so considering dropping these emotional states before performing the face detection will improve the accuracy of the training. The flowchart of the system is illustrated in the Figure 1.

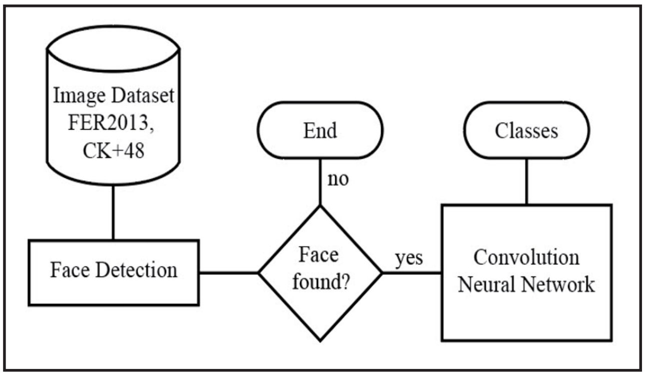

Figure 1. *System flow chart.*

Figure 1 illustrates the process of emotion classification proposed in this paper. First of all, images from the dataset which are the combined dataset between FER2013 and CK+48 with 5 emotional states were fed into the face detection system using the Haar Cascade Classifier [19] to detect face in the input image. If faces are detected from the visual input, the input is passed to ImageDataGenerator function from Keras API for further data augmentation such as flipping, rotating, shearing etc. The face then passes into the CNN classifier to predict classes.

The proposed model used to classify facial expression consists of 4 convolution layers with 32, 64, 128 and 256 filters respectively with the 3x3 kernel. Each layer consist conv3D layer, max-pooling layer, dropout layer, and batch-normalization layer. The conv3D is used to specify convolution kernel, which is 3x3 kernel. The max-pooling layer is used for dimension reduction by finding the maximum value in the 2x2 windows. The dropout layer is included to avoid overfitting problem. The training time optimization is performed by the batch-normalization layer. The activation function using in all layer except the output layer is Exponential Linear Unit (ELU) activation function. The ELU is a function aiming to faster converge cost to zero and boost accurate results. The Categorical Cross-entropy loss is used as a loss function in the output layer. This categorical cross-entropy loss is a combination of the Softmax activation and Cross-entropy loss functions used for multi-class classification to distinguish output into 5 classes of emotion. The overview of the proposed convolution neural network architecture is shown in the Figure 2.

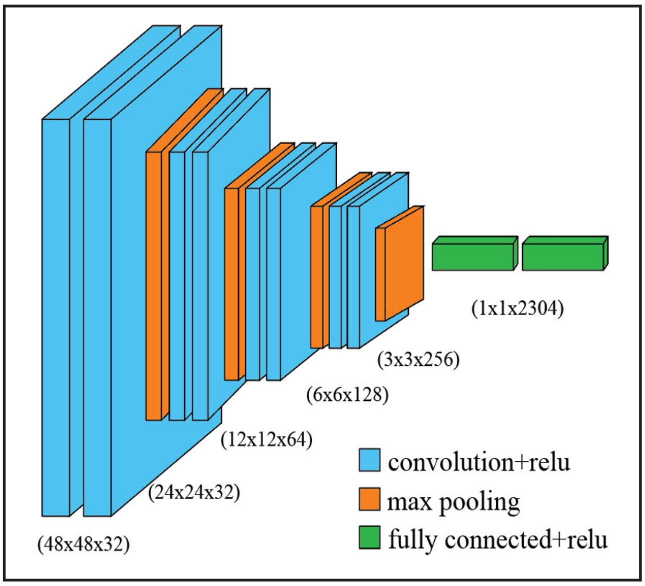

Figure 2. *Architecture of the proposed model.*

4. Model Implementation

4.1 Data Collection and Pre-processing

Datasets used in the experiment has been collected from two sources, which are FER2013 [20] and CK+ [21]. The sample dataset is shown in the Figure 3. These datasets have been preprocessed with 4 main processes including data preparation including disgust and fear dropout, grayscale conversion of RGB images, face detection and crop, image normalization and image augmentation.

4.2 Image Augmentation

A good classifier requires lots of training dataset to achieve a good result. Image augmentation is a process to help to build artificial training images using various methods of image processing. Flipping, rotating, cropping, color jittering, edge enhancement and fancy PCA are popular image augmentation methods. Luke and Geoff [22] have tested these methods for improving deep learning and found that the cropping results in the most accurate classifier followed by rotating and flipping consecutively. In this paper, cropping, rotation, shear, zoom and flip have been used to improve the accuracy of the model.

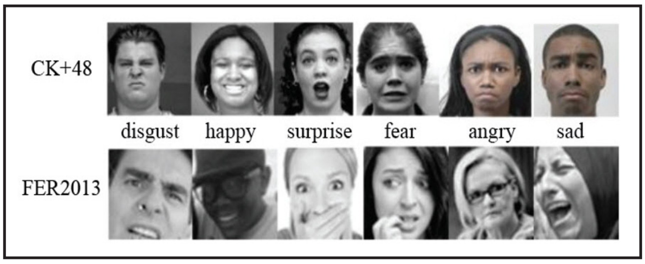

Figure 3. *Samples from CK+48 and FER2013 dataset labeled into emotional states.*

4.3 Model Implementation

The emotion classification model has been written in the python programming language with Keras, Tensorflow [23], NumPy, PIL, OpenCV, and Matplotlib libraries. The Keras provides activation function, optimizers, layers, dropout, batch normalization, etc. The Tensorflow was used as a system backend to accept the inputs of a multidimensional array, which are the pixels of trained images. The OpenCV was used mainly to detect a face in the image or video streaming using Haar cascade classifier [24], grayscale image conversion, and image normalization. The graphic user interface (GUI) was written in a python programming language to accept both still image and real-time video streaming. The system, then, converts the input into 48*48 grayscale image after the face is detected by Haar Cascade Classifier. After that, the cropped image has been passed into the proposed model to classify to 5 distinct emotions.

5. Conclusions

In this research, the main objectives are finding a way to improve emotion recognition system, looking for limitation and proposed model architecture that outweighs the state-of-the-art model. In the experiment, the accuracy and loss assessment of the proposed model against the state-of-the-art model, which is VGG16 in the task of emotion classification based on CK+48 and FER2013 dataset has been performed. 5 emotional states has been selected by dropping out fear and disgust emotional states to improve the model accuracy. The optimized Convolutional Neural Network (CNN) with additional implementation of image augmentations, layer dropout and normalization with less complexity of 1,328,037 parameters has proven to be more efficient compared to the state-of-the-art model with higher complexity of 37,751,765 parameters with the same dataset and numbers of training epochs by achieving a higher rate of model accuracy at 68.28% validation accuracy at 62.87% with the shorter training time compared to 54.27% and 50.24% at 25 epochs as shown in the Figure 4.

***Table 2.** Comparison between model complexities towards accuracy.*

Model	Total parameters	Acc.	Val. Acc.
VGG-16	37,751,765	54.27%	50.24%
Optimized CNN	1,328,037	68.28%	62.87%

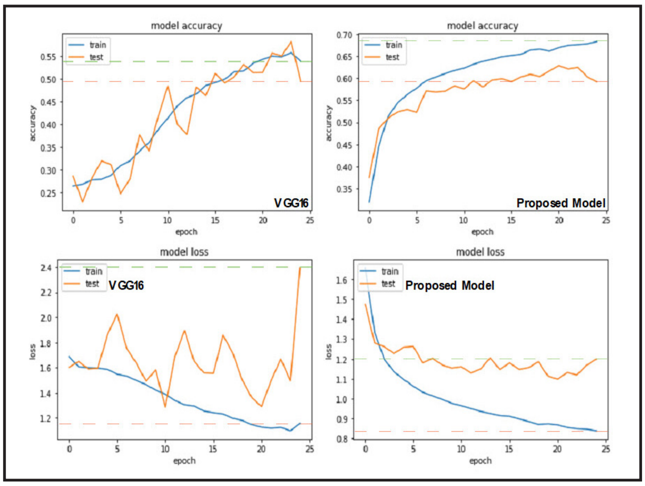

***Figure 4.** Model accuracy and model loss between (left) VGG16 model versus the optimized CNN (right).*

In sum, the proposed model has achieved a worthy result while requiring improvement in some areas e.g. dataset imbalance, lack of dataset, and ambiguity emotion classification between sad and angry emotion in real-time emotion detection. More emotions and features will be included in the near future experiment such as compound emotion classification that can classify multiple emotion at the same time. Moreover, using combined models is also an interesting investigation to improve model accuracy and reduce training time.

6. References

[1] P. Ekman and W. Frieson. *Facial action coding system. Consulting Psychologists Press*, 1977.

[2] J. C. Hagar and et al. *Facial Action Coding System investigator's guide, Salt Lake City, UT: A Human Face*, 2002.

[3] E. Berscheid. "Silent Messages: Implicit Communication of Emotions and Attitudes." *PsycCRITIQUES*, Vol. 26, No. 8, August, 1981.

[4] Very deep convolutional networks for large-scale image recognition. *Visual Geometry Group, Department of Engineering Science*, University of Oxford, 2018.

[5] Yaniv Taigman. "Ming Yang, Marc'Aurelio Ranzato, Lior Wolf. DeepFace: Closing the Gap to Human-Level Performance in Face Verification." *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Columbus, Ohio, pp. 1701-1708, 2014.

[6] A. Ruiz-Garcia, M. Elshaw, A. Altahhan, V. Palade. *Deep Learning for Emotion Recognition in Faces. In:* A. Villa, P. Masulli, A. Pons Rivero (eds) Artificial Neural Networks and Machine Learning – ICANN 2016. ICANN 2016. Lecture Notes in Computer Science, Vol. 9887. Springer, Cham, 2016.

[7] P. Lucey, J. F. 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." *IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops,* San Francisco, CA, pp. 94-101, 2010.

[8] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, D.-H. Lee, Y. Zhou, C. Ramaiah, F. Feng, R. Li, X. Wang, D. Athanasakis, J. Shawe-Taylor, M. Milakov, J. Park, R. Ionescu, M. Popescu, C. Grozea,

J. Bergstra, J. Xie, L. Romaszko, B. Xu, Z. Chuang, and Y. Bengio. "Challenges in representation learning: A report on three machine learning contests." *Neural Networks. Special Issue on, Deep Learning of Representations*, Vol. 64, pp. 59-63, 2015.

[9] P. Tarnowski, M. Kolodziej, A. Majkowski, and R. J. Rak. "Emotion recognition using facial expressions." *International Conference on Computational Science, ICCS 2017*, pp. 1175-1184, June, 2017.

[10] J. Ahlberg. "Using the active appearance algorithm for face and facial feature tracking." *In Proceedings IEEE ICCV Workshop on Recognition, Analysis, and Tracking of Faces and Gestures in Real-Time Systems,* IEEE, pp. 68-72, July, 2001.

[11] 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.

[12] D. H. Kim, M. K. Lee, D. Y. Choi, and B. C. Song. "Multi-modal emotion recognition using semi-supervised learning and multiple neural networks in the wild." *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*. pp. 529-535, November, 2017.

[13] T. L. Nwe, S. W. Foo, L. C. De Silva. Speech emotion recognition using hidden Markov models, *Speech Communication*, Vol. 41, Issue 4, pp. 603-623, 2003.

[14] M. Li and B. Lu. "Emotion classification based on gamma-band EEG." *Annual International Conference of the IEEE Engineering in Medicine and Biology Society*, Minneapolis, MN, DOI: 10.1109IEMBS.2009. 5334139, pp. 1223-1226, 2009.

[15] B. Knyazev, R. Shvetsov, N. Efremova, and A. Kuharenko. *Convolutional neural networks pretrained on large face recognition datasets for emotion classification from video*. arXiv preprint arXiv:1711. 04598, 2017.

[16] V. Vielzeuf, S. Pateux, and F. Jurie. "Temporal multimodal fusion for video emotion classification in the wild." *In Proceedings of the 19th ACM International Conference on Multimodal Interaction*, Glasgow, Scotland pp. 569-576, November, 2017.

[17] S. Robert and W. Adam. "Mouth features extraction for emotion classification." *In 2016 Federated Conference on Computer Science and Information Systems (FedCSIS)*, IEEE, Gdansk, Poland, pp. 1685-1692, September, 2016.

[18] H. Gunes, and M. Piccardi. "Bi-modal emotion recognition from expressive face and body gestures." *Journal of Network and Computer Applications*, Vol. 30, No. 4, pp. 1334-1345, 2007.

[19] L. Cuimei, Q. Zhiliang, J. Nan,and W. Jianhua. "Human face detection algorithm via Haar cascade classifier combined with three additional classifiers." *In 2017 13th IEEE International Conference on Electronic Measurement & Instruments (ICEMI)*, IEEE, Yangzhou, China, pp. 483-487, October, 2017.

[20] P. Giannopoulos, I. Perikos, and I. Hatzilygeroudis. "Deep learning approaches for facial emotion recognition: A case study on fer-2013." *In Advances in Hybridization of Intelligent Methods*. Springer, pp. 1–16, 2018.

[21] P. Lucey, J. F. 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 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*. IEEE, San Francisco, California, pp. 94–101, 2010.

[22] L. Taylor and G. Nitschke. *Improving deep learning using generic data augmentation*. arXiv preprint arXiv:1708.06020, 2017.

[23] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, and M. Kudlur. "Tensorflow: A system for large-scale machine learning." *In 12th {USENIX} symposium on operating systems design and implementation (OSDI),* Savannah, GA, USA, Vol. 16, pp. 265-283, 2016.