EAI Endorsed Transactions
on Pervasive Health and Technology
Research Article EAI.EU

Fusion of Attentional and Traditional Convolutional Networks for Facial Expression Recognition

Tin Trung Nguyen[1,2], Thai Hoang Le[1,2,*]

[1]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam.

[2]Vietnam National University, Ho Chi Minh City, Vietnam.

Abstract

INTRODUCTION: The facial expression classification problem has been performed by many researchers. However, it is still a difficult problem to effectively classify facial expressions in highly challenging datasets. In recent years, the self-weighted Squeeze-and-Excitation block (SE-block) technique has evaluated the importance of each feature map in the Convolutional Neural Networks (CNNs) model, corresponding to the output of the Convolution layer, that has shown high efficiency in many practical applications.
OBJECTIVES: In this paper, with the aim of balancing speed and accuracy for the problem of facial expression classification, we proposed two novel model architectures to solve these problems.
METHODS: Two models proposed in this paper is: (1) a SqueezeNet model combined with a Squeeze-and- Excitation block, (2) SqueezeNet with Complex Bypass combined with a Squeeze-and-Excitation block. These models will have experimented with complex facial expression datasets. Furthermore, the ensemble learning method has also been evidenced to be effective in combining models. Therefore, in order to improve the efficiency of facial expression classification, and aim to compare with the state-of-the-art methods, we use more of the Inception-Resnet V1 model (3). Next, we combine three models (1), (2), and (3) for the classification of facial expressions.
RESULTS: The proposed model gives out high accuracy for datasets: namely, with The Extended Cohn-Kanade (CK+) dataset, there are seven basic types of emotions, reaching 99.10 % (using the last 3 frames), 94.20% for the Oulu-CASIA dataset (from 7th frame) with six basic types of emotions, 74.89% for FER2013.
CONCLUSION: Experimental results on highly challenging data sets (The Extended Cohn-Kanade, FER2013, Oulu-CASIA) show the effectiveness of the technique of combining three models and two proposed models.

Keywords: Facial Expression Recognition, Convolutional Network, Ensemble Learning, Attentional Convolutional Network.

Received on 01 November 2020, accepted on 08 March 2021, published on 17 March 2021

doi: 10.4108/eai.17-3-2021.169033

*Corresponding author. Email: lhthai@fit.hcmus.edu.vn

1. Introduction

Facial expression is characterized by a change in the expression of the face that represents the inner emotional state, thoughts, social communication of an individual [1]. Facial expression recognition is increasingly used in many applications, such as human and machine interaction, driver status monitoring systems. Expression identification is very easy for humans when observing facial expressions via communication intonation. However, the identification of expression by computers is a challenge in computer vision [1]. Recently, facial expression classification has been implemented by many researchers aiming to achieve the same accuracy as a human. It is difficult to separate the feature space of facial expressions, for example, the same expression, each person has different levels of facial expression, resulting in a feature space far away from each other, or two different expression states of the subjects get very close together in a feature space, confusing the classification of expressions. Besides, some expressions such as "happy" and "surprise" look very similar in some cases. Moreover, background states, brightness, pose, all have certain effects on emotional identification.

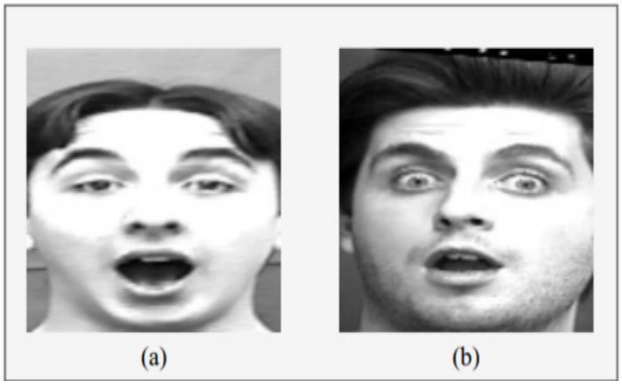

Figure 1. (a) and (b) is same expression surprise but different subject.

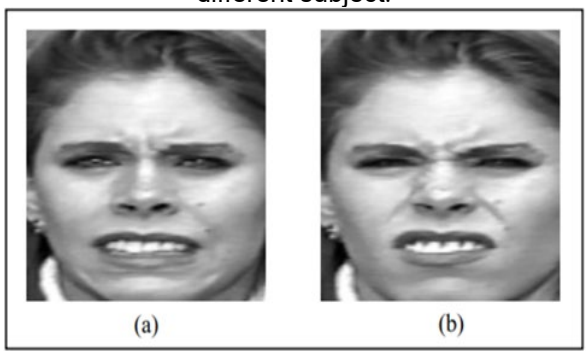

Figure 2. (a) Fear expression and (b) Disgust Expression.

The automatically classifying facial expressions system approached from two different directions. The first one: static-based methods are the system using a single image, extracts features from this single image, and uses the extracted feature to determine which of the seven emotions the static image represents. The second one: dynamics-based methods are the collection of interrelated frames as inputs to the system. Based on temporal information, the frames are combined to extract features for emotional identification. In the classification problem, we care about the discriminative of the extracted feature set. The discrimination is understood that the distance between one class and another class should be the greatest, and the gap between variations within the same class is very small [2]. Accuracy in the classification of facial-based emotions depends on the extraction of features. The better the discrimination is, the higher the accuracy of the classifier performs and vice versa. There are many typical extraction methods available, they can be divided into two categories: (1) Handcrafted feature extraction i.e.: LBP, SIFT, DoG, … (2) Automatic feature extraction, one of the effective techniques for automatic feature extraction is the CNNs. Compared to handcrafted feature extraction methods, the automatic feature extraction by CNNs provides higher accuracy in emotional recognition [3] [4] [5] [6] [7]. In recent years, the CNNs have led to very good performance on a plethora of problems, such as classification, identification, and detection of objects. Particularly in 2015, the CNNs have produced outstanding results in the fields of classification: with dataset Image-Net 1K [8], Resnet152 model has error 3.57%, surpassed humans with error 5% [9]. The input of the CNNs model is a raw image, the CNNs model consists of basic modules such as convolutional layers (Conv), the convolutional kernels are used to create feature map, the subsampling layer helps to retain the desired features, reduces the input size of the feature map to reduce the amount of computation in the convolutional layer. Sub-sampling consists of two main types of max pooling and average pooling. Max pooling, a pooling operation that calculates the maximum value for each patch of each feature map. Average pooling involves calculating the average for each patch of the feature map.

Focus on three data sets CK+, Oulu-CASIA, FER2013, we realize that: although there are different facial expressions for the same emotion, but basically, the facial muscles to represent some emotion are still the same, a feature extractor using CNNs only needs to focus on extracting the features of certain regions in order to distinguish the basic emotions. In CNNs, correlated spatial features are extracted using filters, but features maps extracted from CNNs (feature maps) contain a lot of information that is not useful. In order to minimize the effect of feature maps that are not useful during classification, it's necessary to evaluate the importance of each feature map. Jie Hu et al. [10] proposed the SE-Block technique to solve this issue, and the experiment demonstrated the effectiveness of SE-Block. We decided to study the application of SE-Block in combination with CNNs for facial feature extraction in the facial expression recognition problem from the above knowledge.

Based on those things, we proposed a framework for facial expressions classification based on a deep learning model. Specifically, apply the Multi-task Cascaded Convolutional Networks (MTCNNs) [11] model to face detection, then calibrate bounding box (expand bounding box), combine methods such as image normalization, scale, augmentation (only training), shuffle data (only training) to create inputs for feature extraction and classification phase. In particular, we propose a technique based on combination of three CNNs models belong to the end-to-end networks for expression classification. The proposal model gives out high accuracy for datasets: namely, with The Extended Cohn-Kanade (CK+) [12] dataset, there are seven basic types of emotions, reaching 99.10 % (using the last 3 frames), 94.20% for the Oulu-CASIA [13] dataset (from 7th frame) with six basic types of emotions, 74.89% for FER2013 [14] (seven basic types of emotions). We tested the feasibility of the system on 3 datasets: (1) The Extended Cohn-Kanade (CK+), 7 basic types of emotion dataset including Anger, Contempt, Disgust, Fear, Happiness, Sadness and Surprise, (2) Oulu-CASIA NIR & VIS facial expression database: including 6 emotions Anger, Disgust, Fear, Happiness, Sadness, Surprise, (3) Challenges in Representation Learning Dataset: Facial Expression Recognition Challenge (FER2013) includes 7 types Emotions 7 types of emotions are Angry, Disgust, Fear, Happy, Sad, Surprise, Neutral. In summary, the main contributions of this work are:

- Apply Squeeze-and-Excitation block (SE-block) [10] to create two new models: (1) Model of combining SE-block with Squeezenet [15] (SqueezeNet-SE). (2) Model of combining SE-block with Squeezenet with Complex

Bypass [15] (SqueezeNet Complex-SE). Experiment with 3 dataset sets presented above, both models are highly accurate.

- In order to compare with state-of-the-art methods, we use extra model Inception-Resnet V1 [16], using ensemble learning method combining 3 models: SqueezeNet-SE, SqueezeNet Complex-SE, Inception-ResnetV1.
- Propose a method that combines 3 features from SqueezeNet-SE, SqueezeNet Complex-SE, Inception-Resnet V1 models to classify emotions.
- Experiment the proposal models on three high challenging datasets and all provide high accuracy.

Some of the techniques we use to increase accuracy:

- Initialize the weights from the pre-trained model, which generated from the MS-Celeb-1M dataset [17], loss function is the ArcFace [18] and Softmax Loss [19].
- Use methods such as shuffle data, data augmentation (rotation, random crop, random left-right flip).
- Use the validation method: choose the model with the highest accuracy on the validate set.
- Use ensemble learning to increase accuracy on datasets.
- For data set FER2013 using ten-crop [20] validation method.

The rest of the paper is organized as follows: Section 1: Literature Review, Section 2: Proposal Methods, Section 3: Experiments and Discussions, Section 4: Conclusion.

2. Literature Review

2.1. Related research

Guoying Zhao et al. [21] proposed variations in the local binary pattern (VLBP - Volume Local Binary Pattern) for the extraction of features and classification using SVM (Support Vector Machine). The experimental results were performed on the CK dataset [22] using the cross-validation method (10-fold over the entire frame), with accuracy of 96.26%. Caifeng Shan et al. [23] used LBP (Local Binary Pattern) in combination with Ada-Boost to create Boosted-LBP. Specifically, the most discriminative LBP histograms with Ada-Boost were created for each expression, and then the SVM classifier was used to perform the classification, the experimental results are performed on the CK+ data set, with the cross-validation method (10-fold on the whole frame), the proposed model has 91.4% accuracy. Jie Cai et al. [24] proposed a new error function Island Loss (IL) to enhance the ability to separate features extracted from the CNNs model. In particular, the IL is designed to reduce the variations dimensions of objects in the same class and maximize the distance between one class and the others. Experimented with CK+ dataset, using the last three frames to create 981 images which divided into 10 folds. The authors used the cross-validation method to evaluate with 8 folds for training, 1 fold for validating and 1 fold for testing. Experimental results achieved an accuracy of 98%. This approach is a part of the static-based method. The Yang et al. [25] proposed the De-expression (De-expression Residue Learning) method to classify emotions, Yang et al. used the GANs (Generative Adversarial Networks) [26] model to create a neutral state for each input face image. Yang et al. used the features map from the convolutional layers belonging to both the generator branch and the discriminator, each convolutional layer was passed through the sub-classifier, all sub-classifiers was combined to create the final classifier, the classifier determined the corresponding emotional state for the input image (7 emotional states). Experimented on the CK+ data set, the author used the last 3 frames, used cross-validation to evaluate, divided into 10 frames, Yang et at. achieved 97%. On the Oulu-CASIA data set, the author also used the last 3 frames, using cross-validation for evaluation, divided into 10 frames, achieving 88% accuracy. Kim et al. [27] proposed an approach which combined information from two types of data: alignment (X_A) and non-alignment (X_{NA}), with the purpose of increasing the accuracy of the Facial Expression Recognition (FER) problem. Specifically, for the alignment data: from the original data, the author found the landmark on the face to align face, after that a set of face alignment data (Z_A) was created. Besides, starting with X_A, the author proposed an Alignment-Mapping Networks (AMNs) model to find a face alignment state (Z_A) and a set of feature vectors (h_A). For non-alignment data X_{NA}, this dataset fed into the AMNs and the output is the features vectors (h_{NA}). Next, all X_A, Z_A, X_{NA} was determined by the separate Deep Convolution Neural Networks (DCNs) corresponding to each emotional state. Next, all X_A, Z_A, X_{NA} was fed into Deep Convolution Neural Networks (DCNs) to determine the probability of each emotional layer. In the meantime, h_A and h_{NA} are also fed into the MLP (Multi-Layer Perceptron) networks to determine the probability of each emotional layer. Finally, the Ensemble Learning technique is used, combines at decision level, and labels the emotional state of the input image on the basis of the rules: (1) majority vote, (2) average networks output. The experimental results on the FER2013 data set achieved 73.31% accuracy with both (1) and (2) rules. Isha Talegaonkar et al. [28] proposed a special CNNs architecture to classify emotions. First, the Haar Cascades feature is used for face detection. Second, the normalization technique was used. Finally, the proposed CNNs architecture extracted features and classified the emotional state of the input image set. Experiment on the FER2013 data set, the accuracy on the PublicTest set is 89.78%, the PrivateTest set is 60.12%.

Summary, existing studies: divided into three main groups, (1) handcraft features (2) auto features, and (3) auto features combined with handcraft features. Group (1) precision is not as high as Auto Features, groups (2) and (3), in general, both of these groups use pure CNNs to extract the

feature, without interference with Conv layers. Specifically, using traditional CNNs, do not modify the feature maps through each Conv layer in CNNs, which results in the appearance of many feature maps representing unimportant features in the feature map set created by each Conv layer. Feature maps don't point up important information on areas such as eyes, nose, mouth, while these regions have a significant impact on facial expression recognition. This results in embedded features extracted from CNNs containing only a few important features. Through published studies, we consider these studies focus on post-embedded features extracted from CNNs, they study will process these embedded vectors and suggest improvements. In our approach, focusing on the extraction phase of features improves the discrimination of feature vectors by combining SE-Block with each module (fire-module, fire-module bypass and transition module) included in SqueezeNet.

2.2. SqueezeNet, SqueezeNet with Complex Bypass, Inception Resnet V1

Forrest N. Iandola et al. [15] proposed a SqueezeNet model with a number of parameters less than 50 times when compared to AlexNet [20] but still achieving the same accuracy on the dataset of ImageNet (Large Scale Visual Recognition Challenge 2012 (ILSVRC 2012)) dataset, AlexNet - 57.2%, SqueezeNet reached 57.5%. In addition, the authors also released SqueezeNet with Complex Bypass at the same time, this is a variation of SqueezeNet, derived from the SqueezeNet network, adding a convolutional layer with 1×1 kernel to perform. a skip connection (shortcut connections) [9] in the fire-module. Experimentation has shown that the SqueezeNet with Complex Bypass networks results on ImageNet (ILSVRC 2012) dataset to achieve 58.8 percent accuracy (1.6 percent increase compared to AlexNet). Another CNNs architecture is Inception-Resnet V1 which is made up of a combination of models Inception-A, Inception-B and Inception-C [16]. Skip connection which has been proven to help the model get deeper [9], is also added to each module. The accuracy on 2012 ILSVRC validation set is 78.7%.

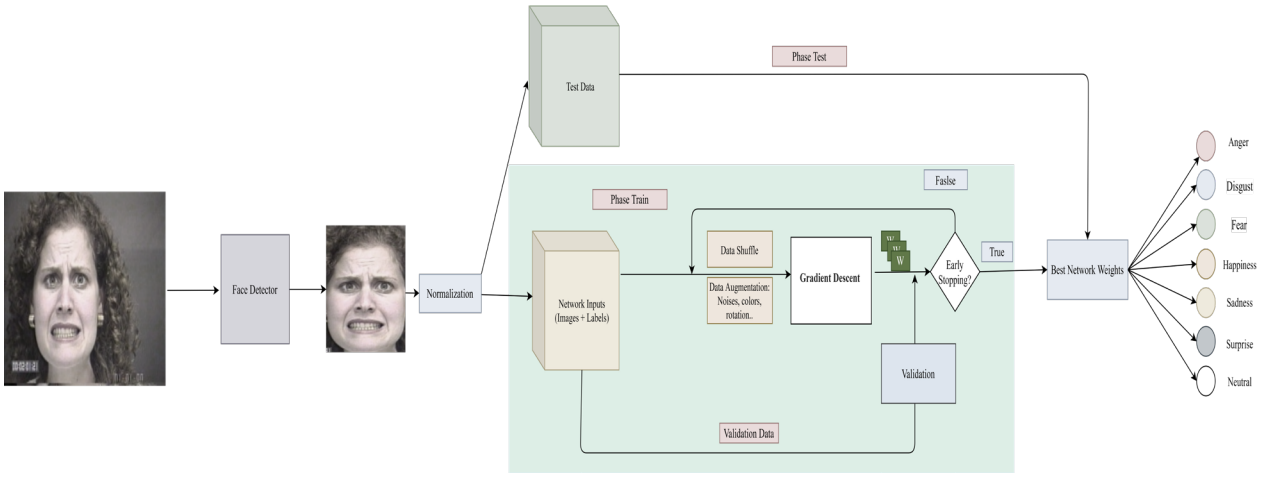

Figure 3. Overview Framework in this work.

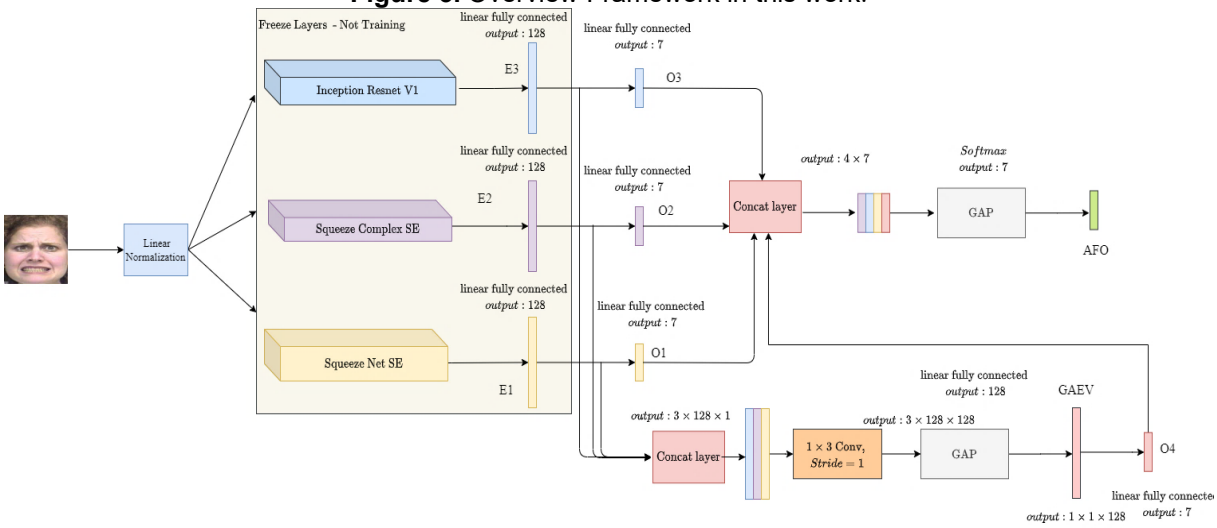

Figure 4. Ensemble model architecture proposed in this work.

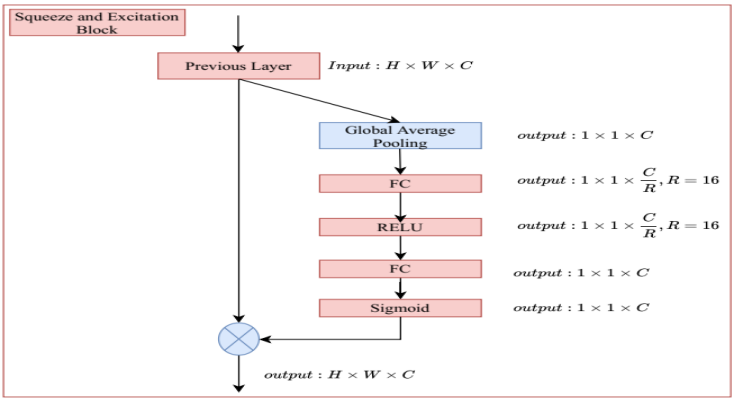

Figure 5. SE-block Architecture, R is the constant used to calculate the number of nodes on the next FC layer (fully connected), C is the number of channels, H is the height of the map features, W is the width of the map features, ⊗ denotes element-wise operator.

2.3. Squeeze and Excitation Block

Squeeze-and-Excitation block [10] (SE-block) (Illustrated in Figure 5) that re-calibrates channel-wise responses by explicitly modelling interdependencies between channels, SE-Net is constructed. And with top-5 error at 2,251%, it won first place in the ILSVRC 2017 [8] classification challenge.

2.4. Pre-trained Model

A pre-trained model is a model that has been trained on another dataset to solve a similar or related problem that we want to solve. MS-Celeb-1M dataset is crawled from the internet, Microsoft Research Public Dataset in 2016, with a number of 10 million images, nearly 100,000 individuals. This dataset is often used to create pre-trained sets for face recognition problems. Chi Jin et al. [29] filtered this dataset again to improve the quality of the dataset (remove overlapping, missing class). We use the dataset was filtered by [29], from which we selected images from subjects with 20-30 samples to train the pre-trained set. Jiankang Deng et al. [18] proposed ArcFace loss function to extract highly discriminative features to solve face verification effectively. Therefore, we used the ArcFace loss function for training CNNs to construct a pre-trained model.

2.5. Ensemble Learning

Ensemble methods is a technique of machine learning which combines several basic models to produce an optimal predictive model. Combining multiple models can increase accuracy in the classification problem. The model combination can be divided into two modes: (1) feature-level (early fusion) (2) decision-level (late fusion). With the combination (1), the features of the sub-models will be merged into a single feature to represent the input and feed into the classifier to assign the label in response to the input. With combination (2), each model makes a decision, the consolidated model will give a final classification based on methods such as (a) voting, (b) average: averaging the probabilities on the output of all models and selecting the class with the highest probability, (c)Weighted: the output of each model will be weighted, showing the contribution of the model. In this paper, we combine the models at both levels: the feature-level, the decision-level. For decision-level, we use the average technique for the classification of facial expression.

3. Proposed Method

In this section, we will discuss the phases of the facial expression classification system of expressions was described in the figure 4.

3.1. Face Detection

At the face detection step, we use the Multi-task Cascaded Convolutional Networks (MTCNNs) method of [11]. MTCNNs is divided into three steps, each of which has a separate CNNs network: P-Net, R-Net, and O-Net. Firstly, the input image is scaled to 5 different ratios to make the input for P-Net. Next, P-Net will return the outputs which are potential faces regions. These regions will be adjusted (padding), then scaled to 24×24 pixels to become input for the next R-Net network. R-Net Networks removes non-face regions, calibrates region coordinates using bounding box regression. The output of R-Net is similar to P-Net: potential face regions. Calibration technique (padding, Non-maximum Suppression - NMS) is used for these areas. Next, the potential face regions are scaled to 48×48 pixels and fed into the O-Net. O-Net will again classify regions with faces and not faces. For regions with faces, the O-Net returns the confidence score and the bounding box coordinates are adjusted. Finally, the NMS is used to calculate the bounding box coordinates for each

face found. After we have found a face in image, we increase the bounding box coordinates for each face to 20 pixels (for both width and height). Finally, each face region will be cropped and resized to 128×128 pixels.

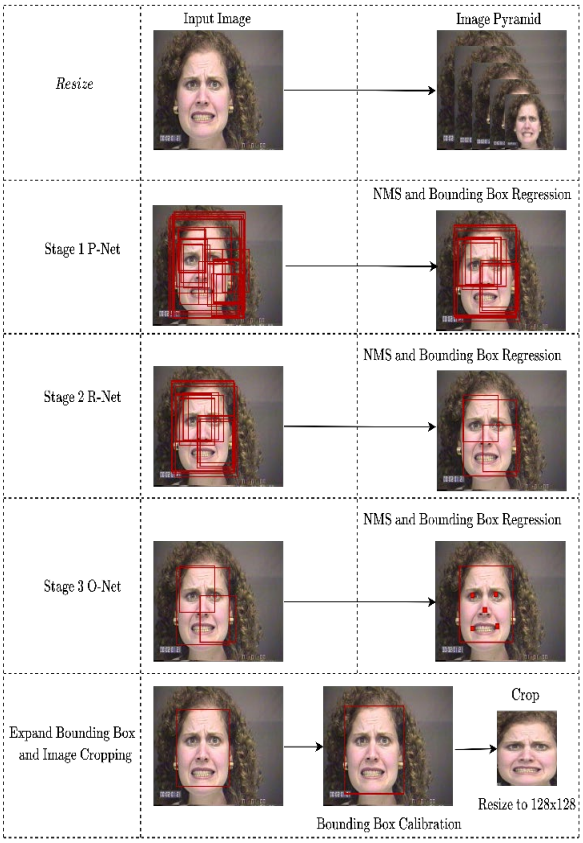

Figure 6. Illustrate all stage of MTCNN.

3.2. Down-Sampling

With the aim of using Inception-Resnet V1 effectively, the FER2013 image faces, with an original size of 48×48 pixels, will be resized to 86×86 pixels. For both Oulu-CASIA, CK+ dataset, we resized them to 128×128 pixels after face detection.

3.3. Augmentation

Affine transformations and other transformations often used to generate more data in deep learning such as rotation, scaling, translation, black-out, random crop, ten-crop [17], and brightness and colours transformations. Data augmentation helps avoid overfitting in deep learning [28]. In this paper, we use the methods: random rotation with the value angle in the domain [-15, +15], random vertical flipping, random contrast.

3.4. Normalization

We apply the linear transformation method mentioned by Bishop in [30] to normalize the input image for both the train and the test phase. The linear transformation of the pixel value of the input images is one of the common and simple forms of pre-processing for input normalization. The linear transformation used in this paper will bring all the original values to the same smaller range. This linear transformation ensures effective input normalization for images with property: (1) Images of the same subject but with different contrasts, (2) The pixel values of the image are varied (different lighting conditions). The input linear transformation method is carried out through the following steps:

- Step 1 for each image, calculate the mean value $\bar{x}$ of the image and variance value σ^2

$$\bar{x} = \frac{1}{\mathrm{W} \times H} \sum_{i=1}^{H} \sum_{j=1}^{\mathrm{W}} x_{ij} \quad (1)$$

$$\sigma^2 = \frac{1}{\mathrm{W} \times \mathrm{H}} \sum_{i=1}^{H} \sum_{j=1}^{\mathrm{W}} (x_{ij} - \bar{x})^2 \quad (2)$$

Where W is the width, H is the height of the input image x, x_{ij} is the pixel value with the coordinates (i, j), $\bar{x}$ is mean value of image, σ^2 is variance.

- Step 2 apply linear transformation for the image is calculated as follows

$$\tilde{x} = \frac{x - \bar{x}}{|\sigma|} \quad (3)$$

3.5. CNNs architectures were proposed in this work

In order to effectively solve the problem of image classification on ImageNet [31]. Iandola et al. [15] proposed two architectures: SqueezeNet and SqueezeNet with Complex Bypass. SqueezeNet model with a number of parameters less than 50 times when compared to AlexNet but still achieve the same accuracy on ImageNet (ILSVRC 2012) dataset. Specifically, SqueezeNet included 8 blocks of fire-module and layers: max-pooling, Global Average Pooling (GAP), and Fully Connected (FC). Each fire-module consists of two consecutive layers: the squeeze layer and the expand layer. In which the squeeze layer is a Convolutional layer with a kernel size was 1×1, it reduces the number of feature maps, and then feeds the output to the expand layer. Expand layer formed from Convolutional layer mixes with 1×1 kernel size and 3×3 kernel size Convolutional layer. SqueezeNet with Complex Bypass similar to SqueezeNet, with a few changes: the skip connection is added to the fire-module, while the 1x1 kernel-sized Convolutional layer is added to some fire-modules to make the transition module. The purpose of the transition module is to adjust the number of fire-module feature maps on the previous layer, so that it was equal to the number of current fire-module feature maps to perform

the skip connection. In this paper, we want to leverage the high performance of two above models to propose two extended models of SqueezeNet and SqueezeNet with Complex Bypass, called: (1) SqueezeNet-SE, and (2) SqueezeNet Complex-SE, in order to effectively solve the problem of facial expression classification. Model (1) (SqueezeNet-SE) consists of 9 fire-module blocks, and each fire-module is combined with SE-block (Figure 8 describes the proposed SqueezeNet-SE model). Figure 7 shows the combination of the fire module and the SE-block in the proposed model. Applying SE-block to the expand layer output is to recalibrate feature maps to highlight important feature maps. Thus, corresponding to 9 fire-module blocks, 9 SE-blocks will be used to recalibrate feature maps. Model (2) (SqueezeNet Complex-SE) is illustrated in Figure 9. Figure k illustrates the operation of the fire-module bypass, includes the steps: (a) the execution of the skip connection; (b) the application of the SE-block. The Transition Module is a fire-module with an additional 1×1 kernel size Convolutional layer, as illustrated in Figure 2. Similar to SqueezeNet-SE, after the skip connection is performed, SE-block is used to recalibrate feature maps to indicate the importance of feature maps feature maps in each module (both Fire-module bypass (figure 10) and Transition-module (figure 11)). Activation Rectified Linear Units [32] (ReLU) is applied after each layer Conv.

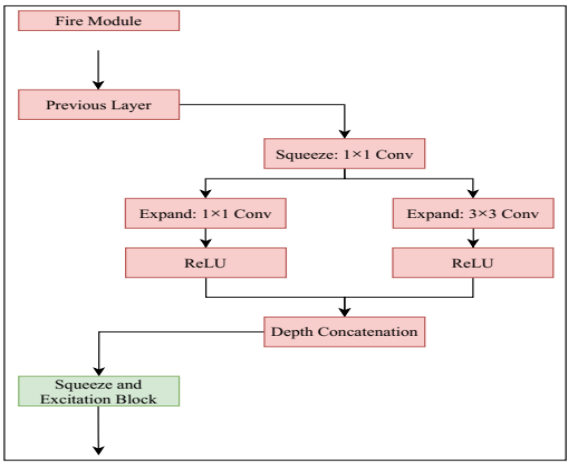

Figure 7. Fire Module.

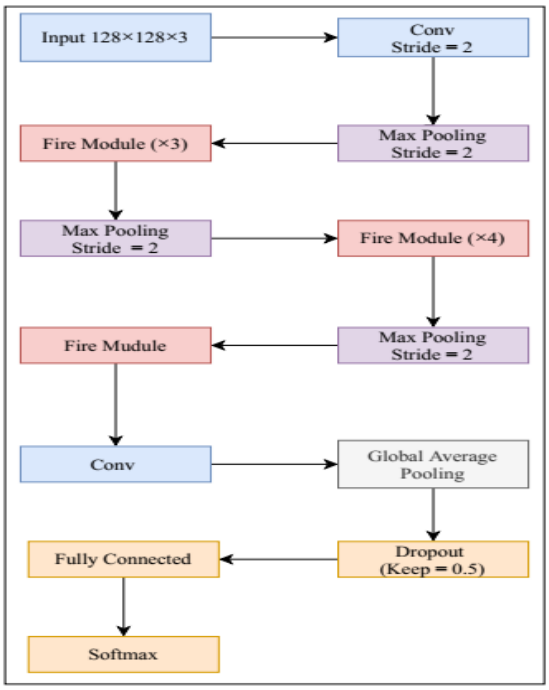

Figure 8. SqueezeNet SE in this work.

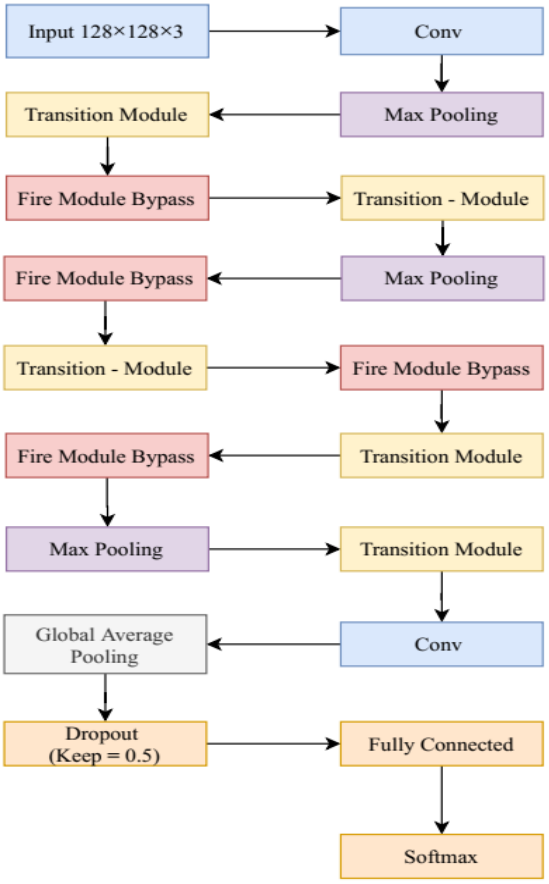

Figure 9. SqueezeNet Complex-SE in this work.

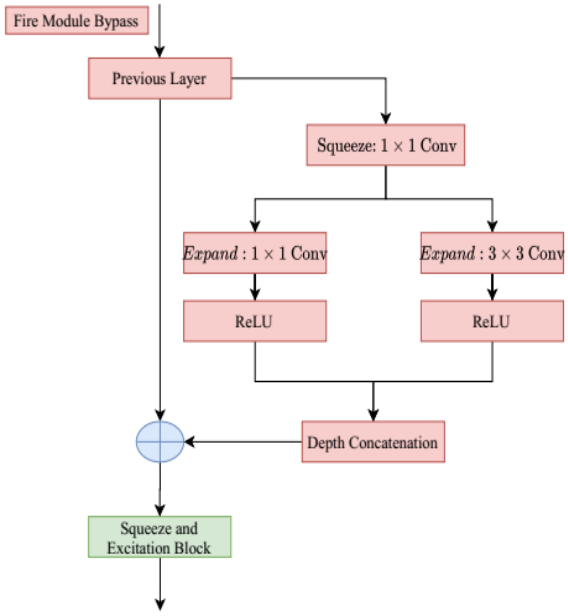

Figure 10. Fire-module Bypass, ⊕ denotes element-wise addition.

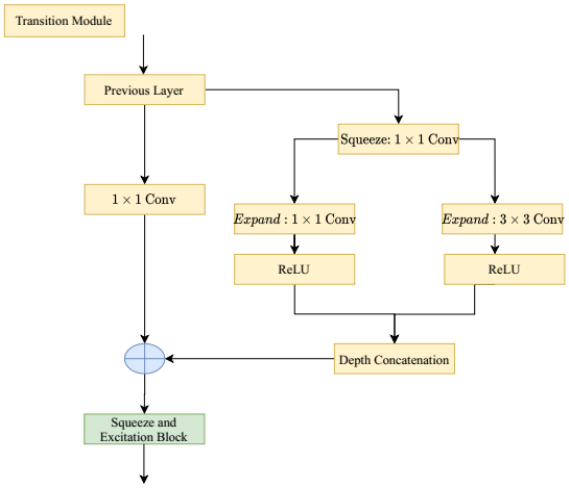

Figure 11. Transition module, ⊕ denotes element-wise addition operator.

3.6. Ensemble for facial expression recognition

In this period, we combining the models together. The output of the SqueezeNet-SE model is an embedding vector of 128 dimensions, named E1. E1 will be fed into a Fully Connected (FC) that generates O1 output, with an output size of 7 nodes, corresponding to 7 facial emotions. Similar to SqueezeNet-SE, the output of SqueezeNet Complex-SE is an embedding vector called E2, 128-dimensional. E2 is also used as input for the FC that generates O2 (the size is 7 nodes). Finally, in order to increase the efficiency of the facial emotions classification system, we use more of the Inception-Resnet V1 model, with the architecture proposed in [13]. The output of the Inception-Resnet V1 model in this paper is an embedding vector called E3, 128-dimention. E3 will be fed to a FC, the output of this FC is called O3 with a size of 7 nodes. We combine 3 vectors E1, E2, E3 together to create a 128×3 matrix. Next, use a 1×3 kernel Convolutional layer to synthesize information from 3 vectors, create feature maps (new feature space is 128×3×128). Using the Average Pooling Operator creates a new feature space named AEV (Average Embedding Vectors), which has a 128×1×1 dimension. AEV will be fed into an FC layer to generate O4, with 7 nodes in size. Combine O1, O2, O3, O4 and form a 7×4 matrix. Using an average calculation for 7×4 matrix, get a vector called AFO (Average Final Output), with a size of 7×1, which corresponds to 7 emotions.

4. Experiments and Discussions

4.1. Datasets

The system is tested and trained on three databases: FER2013, The Extended Cohn-Kanade (CK+), Oulu-CASIA.

CK+: The Extended Cohn-Kanade (CK+) [12] dataset was published in 2010 by Cohn-Kanade et al. based on the CK dataset. CK+ contains 593 sequences, belonging to 123 subjects, but there are only 327 sequences of 118 subjects labelled with seven basic expressions: anger, contempt, disgust, fear, happiness, sadness and surprise. Figure 12 illustrate some of images from the CK+ dataset.

Figure 12. Samples in CK+ Dataset.

Oulu-CASIA [13]: includes 2880 sequences, consists of six expressions (Surprise, Happiness, Sadness, Anger, Fear and Disgust) from 80 people between the ages of 23 and 58. Each of the sequences is captured with one of two imaging systems: (1) Near-infrared (NIR), (2) Visible light (VIS). Each imaging system is made under three different lighting conditions: Dark, Strong, Weak. Figure 13 illustrate some of images from the Oulu-CASIA dataset.

Figure 13. Samples in Oulu-CASIA dataset.

FER2013: FER2013 dataset published during the ICML competition in 2013 [14]. Dataset FER2013 is a large dataset with images of human faces showing emotions. This dataset was collected from the internet in the following context: reality, freestyle, non-constraint, all images 48×48 pixels in size. This data set is divided into three subsets: (1) training set: 28,709 images (2) public test set (validation set): 3,589 images, (3) private test set (test set): 3,589 images. The FER2013 dataset includes 7 labels: Anger, Disgust, Fear, Happiness, Sadness, Surprise and Neutral. Figure 14 illustrates some of images from the FER2013 dataset.

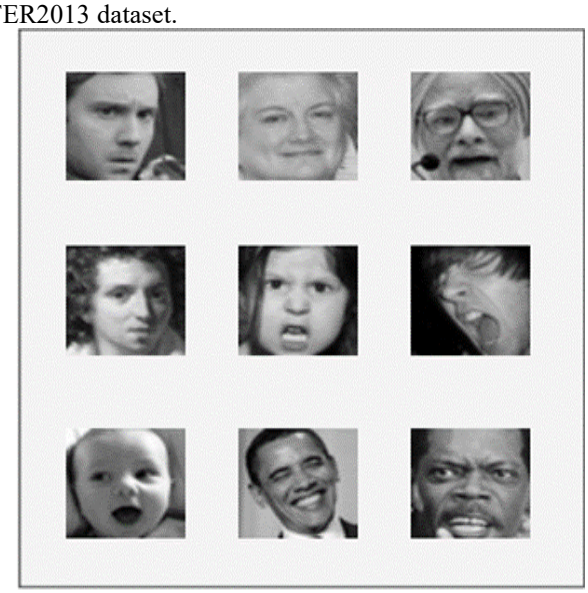

Figure 14. Samples in FER2013 dataset.

Summary, for CK+ and Oulu-CASIA dataset: Sequence data, lab context, video sequence starts with a neutral state. Expression level increases over time, peaking at the last frame, and the databases are all labelled as basic emotions, the common challenge of these two datasets is that the amount of data is too small compared to the datasets tested using the CNN method. The amount of data is too small, leading to low generalization. The data of two datasets are a variety: male-female, skin colour, light levels vary between videos (for CK+). It is possible to have expressions of relatively different facial muscle groups with the same emotion, both male and female. This leads to confusion in the classification of emotions. For FER2013: The challenge of the FER2013 dataset: Collected from the internet, static images, free context, different light levels (difference grayscale level), face with free-angle. Distribution at all ages, the image was incorrectly labelled, the image did not contain a human face. This dataset also contains faces with glasses and hats. Moreover, with the same emotion, objects of different ages show some difference.

4.2. Experiments

Based on the models proposed in section 3: (1) Inception-Resnet V1, (2) SqueezeNet-SE, (3) SqueezeNet Complex-SE, (4) ensemble model, we implement an Emotion Recognition System for each model. At the same time, 3 datasets FER2013, CK+, Oulu-CASIA were used to experiment, evaluate and analyse the feasibility of the system implemented. Implementation details.

- CK+: The CK+ dataset is not divided into train, test, and validate set as FER2013, so we extract the last frames (last 3 frames) for each sequence. After that, the cross-validation method is applied, the images are further divided into 10 folds in ascending order, based on identity (9 folds for training, 1 fold for testing). The validation set is split from the training data (10% of the number of samples in 9 folds).
- Oulu-CASIA: We use images from the VIS system, with strong lighting (Strong) from frame 7 to the end. A 10-fold subject-independent cross-validation is performed, as with the CK+ experimental setting. The validation set is split from the training data (15% of the number of samples in 9 folds).
- FER2013: the images were resized to 96×96 pixels, then the 10-crop method was used to generate data for three training sets, Public Test, Private Test. The size of the cropped image is 86×86 pixels. The original image is also resized to 86×86 pixels at the end of the ten-crop implementation, forming 11 images, using the voting method to select the final classification for the input image (for the validation and test set).
- Generating pre-trained model: with 2 proposed models and Inception–Resnet V1, generating three pre-trained models respectively, the pre-trained generation process is referred to in Section 2.4.

- Training parameters for 4 models: we use a Stochastic Gradient Descent (SGD) optimizer for both the creation of a pre-trained model and the fire-tuning. Learning rate Schedule is used, change the learning rate based on the accuracy of the validation set. Weight-decay = 5e-5, dropout rate = 0.5, batch size = 96, random seed = 777, The loss function used during FER training is Softmax Loss, Softmax Loss is actually just Softmax Activation plus Cross-Entropy Loss [19]. Images are represented in RGB colour space.
- Number of parameters: Inception-ResnetV1: 22.8 M (Million), SqueezeNet Complex-SE: 4.4M, SqueezeNet-SE: 2.6M.
- Metrics: use accuracy metric [33] and confusion matrix [34] to evaluate the proposed models.
- Early Stopping: we keep the model with the highest accuracy on the validation set as the final model (figure 15: illustration of Early Stopping technique for dataset FER2013).
- The environment-tested system: Ubuntu 18.04 LTS, Tensorflow 1.9, CUDA 9.0, GPU Titan Xp 12 GB memory, 64 GB RAM, Corei7-5820 K Intel 3.30GHz, 12 cores.

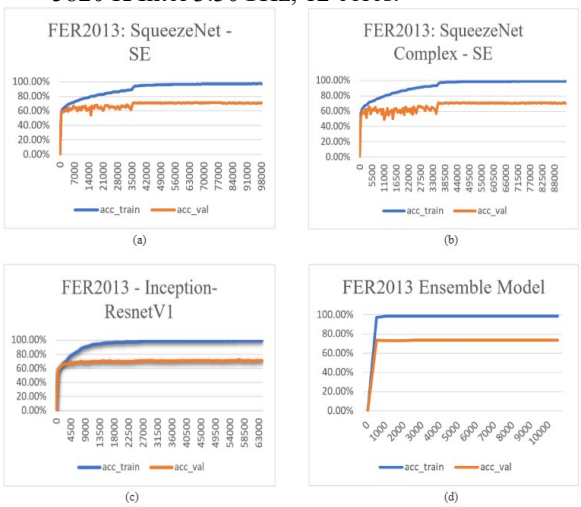

Figure 15. Accuracy on Training Set and Validation Set (acc_train: accuracy training set, acc_val: accuracy validation set).

4.3. Results

The tables 1 and 2 show the accuracy of the classification (%) for each emotion in two datasets: CK+ and Oulu-CASIA.

Table 1. Accuracy (%) per expression on dataset CK+.

Model	Anger	Contempt	Disgust	Fear	Happy	Sad	Surprise
Inception-Resnet V1	98.52	100	100	100	100	100	100
Squeezenet-SE	100	100	100	100	100	100	99.6
Squeezenet Complex-SE	99.29	89	100	100	100	100	99.6
Ensemble Model	99.29	91	100	100	100	100	100

Table 2. Accuracy (%) per expression on dataset Oulu-CASIA.

Model	Anger	Disgust	Fear	happiness	Sadness	Surprise
Inception-Resnet V1	90.35	94.95	91.23	96.06	94.06	96.15
SqueezeNet -SE	94.06	86.09	98.08	97.65	92.33	88.38
SqueezeNet Complex-SE	88.98	89.62	96.06	94	83.49	85.52
Ensemble Model	94.93	91.63	98.06	96.87	92.33	90.57

For the table 1: all three proposed models (SqueezeNet - SE, SqueezeNet Complex-SE, ensemble model) all showed positive results, with the lowest accuracy of 89% for contempt expression belongs to the SqueezeNet Complex-SE model, and 100% for other emotions.

For the table 2: all three proposed models are tested on Oulu-CASIA give quite good results. The lowest accuracy of 83.49% for sad expressions belongs to SqueezeNet Complex-SE model. The highest accuracy is 98.6% for fear expression belongs to ensemble model.

Table 3 shows accuracy on the CK+ dataset for 7 expressions. Accuracy for 6 expressions on the Oulu-CASIA dataset is displayed in table 4. Table 5 shows lists the accuracy of 7 expressions for the FER2013 dataset.

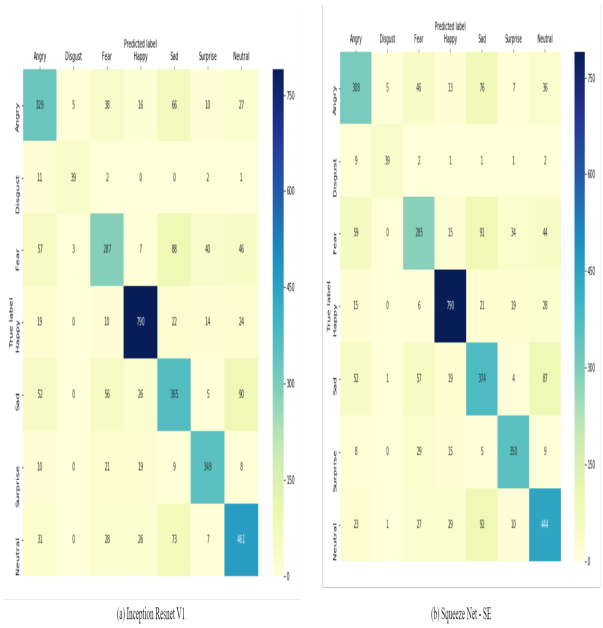

Figure 16. Confusion Matrix for FER2013: (a) Inception-ResnetV1 (b) SqueezeNet SE.

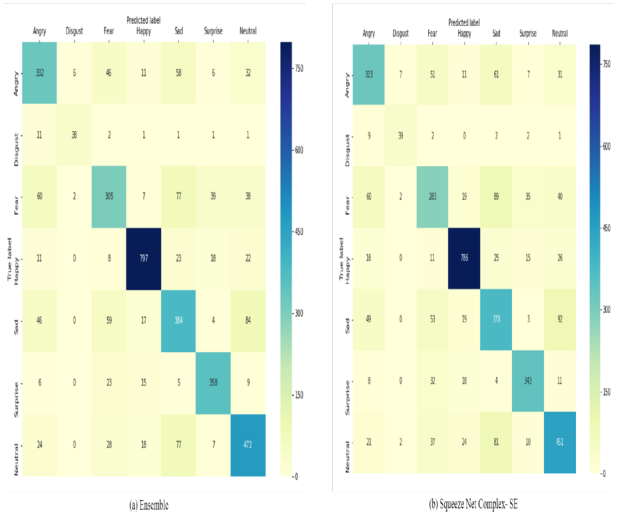

Figure 17. Confusion Matrix for FER2013: (a) SqueezeNet Complex-SE (b) Ensemble Model.

Table 3. Results of 7 expressions on the CK+ dataset.

Studies	Method	Accuracy (%)
Liu et al. [35] (2017)	CNNs (last three frames – static based)	97.1
Cai et al. [24] (2017)	CNNs (last three frames – static based)	94.39
Meng et al. [36] (2017)	CNNs (last three frames – static based)	95.37
Yang et al. [25] (2018)	cGAN (last three frames – static based)	97.30
Inception-Resnet V1 (Ours)	CNNs (last three frames – static based)	99.80
SqueezeNet-SE (Ours)	CNNs (last three frames – static based)	99.50
SqueezeNet Complex-SE (Ours)	CNNs (last three frames – static based)	98.3
Ensemble Model (Ours)	Ensemble CNNs (last three frames – static based)	99.1

The results of three proposed models on three datasets are comparable with other authors' methods (see tables 3, 4, 5). Specifically, with the FER2013 dataset, the ensemble model (74.87%) has 0.55% lower accuracy than the [35] model (75.42%), and 1-3% higher accuracy than other models (table 5). Similarly, three proposed models have 2-4% higher results on the CK+ dataset than those in other authors' studies (table 3). Observing table 4, three proposed models also have higher results than the models of other authors on the Oulu-CASIA dataset.

Table 4. Results of 6 expressions on the Oulu-CASIA dataset. (DB: dynamics-based, SB: Static-based, En-CNNs: Ensemble CNNs)

Studies	Method	Accuracy (%)
Zhao et al. [37] (2016)	CNNs (DB)	84.59
Zhang et al [38]. (2017)	En-CNNs(DB)	86.25
kuo et al [39]. (2018)	En-CNNs(DB)	91.67
Yu et al [40]. (2017)	CNNs (DB)	86.23
Inception-Resnet V1 (Ours)	CNNs (SB)	93.90
SqueezeNet-SE (Ours)	CNNs (SB)	92.80
SqueezeNet Complex-SE (Ours)	CNNs (SB)	89.7
Ensemble Model (Ours)	En-CNNs(SB)	94.20

Table 5. Results of 7 expressions on the FER2013 dataset.

Studies	Method	Accuracy (%)
Hua et al. [41] (2019)	Ensemble CNNs	71.91
Georgescu et al. [42] (2019)	Ensemble CNNs	75.42
Connie et al [43]. (2017)	Ensemble CNNs + SIFT	73.40
Li et al [44]. (2017)	Ensemble CNNs	70.66
Kim et al. [45] (2016)	Ensemble CNNs	72.72
Xu et al. [46] (2018)	CNNs	69.70
Minaee et al. [47] (2019)	CNNs	70.02
Inception-Resnet V1 (Ours)	CNNs	73.00
SqueezeNet-SE (Ours)	CNNs	72.16
SqueezeNet Complex-SE (Ours)	CNNs	72.53
Ensemble Model (Ours)	Ensemble CNNs	74.87

5. Discussion

Analysing the problem of classifying facial emotions, researchers have shown that facial features have different

effects for effective emotional classification, specifically, regions such as eyes, nose, and mouth have a higher effect on the performance of expression classification than other regions [47]. Thus, the use of the SE-block in combination with CNNs to re-calibrate feature maps (feature maps are weighted) is necessary for the problem of facial expression classification. Thanks to the SE-Block calibration feature maps, important features are retained and the influence of unimportant features is reduced immediately after each module in CNNs. This allows for the extraction of embedded vectors by CNNs with highly discriminative features, resulting in improved performance of layering in the classifier. The values of the important feature maps will be changed less (weight close to 1.0) and the non-important feature maps will be suppressed (weight close to 0.0). Therefore, we proposed two models for combining SE-block and CNNs: (1) SqueezeNet-SE is a combination of SE block and SqueezeNet; (2) SqueezeNet Complex-SE is a combination of SE-block and SqueezeNet with Complex Bypass.

To more clearly, in [9], the Facial Action Code System (FACS) was described by Patrick Lucey et al., this system describes a set of facial muscle movements that correspond to a displayed emotion on the face. From the information on the FACS system, realize that: action units that define basic emotions are mainly muscle groups in areas around the eyes, nose, and mouth. It is necessary to focus on extracting features from regions around these locations for high accuracy in facial expression recognition. For more effective facial expression recognition, information from the regions (eyes, nose, and mouth) must be extracted from CNNs. However, with the number of feature maps generated by each Conv layer in CNNs, a lot of weak information areas (not nearly three areas of the eyes, nose, and mouth) will also participate in the classification process, thus it is necessary to minimize the effect of a feature map containing unimportant information, Since then it has been necessary to apply SE-Block (a way to automatically recalibrate feature maps) to automatically minimize the impact of unimportant feature maps and to retain only important feature maps.

Final, our model has higher results on the two datasets CK+ and Oulu-CASIA than the current models (Table l). Our method gives results on the FER2013 dataset with an accuracy close to model [42] (accuracy 0.56 less than model [42]), however our model is simpler. In addition, Cires¸an et al. [48] has shown that combining the models together achieves higher classification accuracy than using a single model. Details, Cires¸an et al. [48] used the ensemble model to solve the problem of image classification, experimented with six datasets (MNIST, NIST SD 19, HWDB1.0 (on – off), CIFAR10, traffic signs, NORB). Experimental results show that the ensemble model is more efficient than the single model. We also use ensemble model in this paper (Section 3.6). Experiment shows that ensemble model has better accuracy than single model (on two datasets: FER2013, Oulu-CASIA) and equivalent accuracy on CK+ dataset.

6. Conclusion

In this paper, we proposed 2 models SqueezeNet-SE and SqueezeNet Complex-SE, which is a combination of CNNs and SE-block. Furthermore, a model combining the two proposed models and the Inception-Resnet V1 model (Ensemble model) is also proposed. The proposed models are experimentally evaluated on three complex and challenging datasets: FER2013, Oulu-CASIA, CK+. The experimental results show the feasibility of the proposed models.

In the future, we will: (1) In the case of video data, it is necessary to develop a method for using more relationships between frames over time. (2) Develop, expand, and test the efficiency of the system with other classification problems. (3) Combine both Attention Spatial and Attention Channels in order to enhance the ability to extract features.

Acknowledgements.
This research is funded by Vietnam National University, Ho Chi Minh City (VNU-HCMC) under grant number C2020-18-18.

References

[1] S. Z. Li and A. K. Jain, Handbook of Face Recognition, London: Springer, 2011.

[2] Y. Wen, K. Zhang, Z. Li and Y. Qiao, “A Discriminative Feature Learning Approach for Deep Face Recognition,” *ECCV 2016: Computer Vision – ECCV 2016,* vol. 9911, pp. 499-515, 2016.

[3] J. Anil and L. P. Suresh, "Literature survey on face and face expression recognition," *International Conference on Circuit, Power and Computing Technologies (ICCPCT),* pp. 1-6, 2016.

[4] S. Li and W. Deng, “Deep Facial Expression Recognition: A Survey,” *IEEE Transactions on Affective Computing,* 2020.

[5] V. . M. Duc, A. Sugimoto and L. H. Thai, “Facial Expression Recognition by Re-ranking with Global and Local Generic Features,” *2016 23rd International Conference on Pattern Recognition (ICPR),* 2016.

[6] V. D. Minh and L. H. Thai, “Deep Generic Features and SVM for Facial Expression Recognition,” *2016 3rd National Foundation for Science and Technology Development Conference on Information and Computer Science,* 2016.

[7] K. Wang, X. Peng, J. Yang, D. Meng and Y. Qiao, "Region Attention Networks for Pose and Occlusion Robust Facial Expression Recognition," *: IEEE Transactions on Image Processing,* vol. 29, pp. 4057 - 4069, 2020.

[8] O. Russakovsky,, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg and L. Fei-Fei, "ImageNet Large Scale Visual Recognition Challenge," *International Journal of Computer Vision, 2015.*

[9] K. He, X. Zhang and S. Ren, "Deep Residual Learning for Image Recognition," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 770-778, 2016.

[10] J. Hu, L. Shen, S. Albanie, G. Sun and E. Wu, "Squeeze-and-Excitation Networks," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition,* pp. 7132-7141, 2018.

[11] K. Zhang, Z. Zhang, Z. Li and Y. Qiao, "Joint Face Detection and Alignment using Multi-task Cascaded Convolutional Networks," *IEEE Signal Processing Letters,* vol. 23, pp. 1499-1503, 2016.

[12] 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," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops,* pp. 94-101, 2010.

[13] M. Taini, G. Zhao, S. Z. Li and M. Pietikainen, "Facial expression recognition from near-infrared video sequences," *2008 19th International Conference on Pattern Recognition,* pp. 1-4, 2008.

[14] I. J. Goodfellow, D. Erhan, P. L. Carrier, A. Courville, M. Mirza, B. Hamner, W. Cukierski, Y. Tang, D. Thaler, L. Dong-Hyun, Y. Zhou, C. Ramaiah, F. Fangxiang, 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 Information Processing. ICONIP 2013. Lecture Notes in Computer Science,* vol. 8228, 2013.

[15] F. N. Iandola, S. Han, M. W. Moskewicz, K. Ashraf, W. J. Dally and K. Keutzer, "SqueezeNet: AlexNet-level accuracy with 50x fewer parameters and <0.5MB model size," *arXiv preprint arXiv:1602.07360,* 2016.

[16] C. Szegedy, S. Ioffe, V. Vanhoucke and A. Alemi, "Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning," *AAAI Conference on Artificial Intelligence,* 2016.

[17] Y. Guo, L. Zhang, Y. Hu, X. He and J. Gao, "MS-Celeb-1M: A Dataset and Benchmark for Large-Scale Face Recognition," in *Computer Vision – ECCV 2016*, Springer International Publishing, 2016.

[18] J. Deng, J. Guo, N. Xue and S. Zafeiriou, "ArcFace: Additive Angular Margin Loss for Deep Face Recognition," *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),* pp. 4685-4694, 2019.

[19] I. J. Goodfellow, Y. Bengio and A. Courville, "6.2.2.3 Softmax Units for Multinoulli Output Distributions," in *Deep Learning*, MIT Press, 2016, p. 180–184.

[20] A. Krizhevsky, I. Sutskever and G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," *Neural Information Processing Systems,* vol. 25, 2012.

[21] G. Zhao and M. Pietikäinen, "Dynamic Texture Recognition Using Local Binary Patterns with an Application to Facial Expressions," *IEEE transactions on pattern analysis and machine intelligence,* vol. 29, pp. 915-28, 2007.

[22] P. Lucey, J. F. Cohn and T. Kanade, "Comprehensive database for facial expression analysis," *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580),* pp. 46-53, 2000.

[23] C. Shan, S. Gong and P. W. McOwan, "Facial expression recognition based on Local Binary Patterns: A comprehensive study," *Image and Vision Computing,* vol. 27, pp. 803-816, 2009.

[24] J. Cai, Z. Meng, A. S. Khan, Z. Li, J. O'Reilly and Y. Tong, "Island Loss for Learning Discriminative Features in Facial Expression Recognition," *2018 13th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2018),* pp. 302-309, 2018.

[25] H. Yang, U. Ciftci and L. Yin, "Facial Expression Recognition by De-expression Residue Learning," in *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2018, pp. 2168-2177.

[26] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville and Y. Bengio, "Generative Adversarial Nets," in *Advances in Neural Information Processing Systems 27*, Curran Associates, Inc., 2014, pp. 2672-2680.

[27] B.-K. Kim, S.-Y. Dong, J. Roh, G. Kim and S.-Y. Lee, "Fusing Aligned and Non-aligned Face Information for Automatic Affect Recognition in the Wild: A Deep Learning Approach," in *2016 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2016, pp. 1499-1508.

[28] I. Talegaonkar, K. Joshi, S. Valunj, R. Kohok and A. Kulkarni, "Real Time Facial Expression Recognition using Deep Learning," *Proceedings of International*

Conference on Communication and Information Processing (ICCIP) 2019, 2019.

[29] C. Jin, R. Jin, K. Chen and Y. Dou, "A Community Detection Approach to Cleaning Extremely Large Face Database," *omputational Intelligence and Neuroscience,* 2018.

[30] C. M. Bishop, "8.2 Input Normalization and encoding," in *Neural Networks for Pattern Recognition*, 198 Madison Ave. New York, NYUnited States, Oxford University Press, Inc, 1996.

[31] J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li and L. Fei-Fei, "ImageNet: A Large-Scale Hierarchical Image Database," in *CVPR09*, 2009.

[32] F. A. Agarap, "Deep Learning using Rectified Linear Units (ReLU)," *ArXiv,* vol. abs/1803.08375, 2018.

[33] D. Powers, "Evaluation: From Precision, Recall and F-Factor to ROC, Informedness, Markedness & Correlation," *Mach. Learn. Technol.,* vol. 2, 2008.

[34] K. M. Ting, "Confusion Matrix," in *Encyclopedia of Machine Learning*, Boston, MA, Springer, 2011, p. 209.

[35] X. Liu, J. You, B. Vijaya Kumar and P. Jia, "Adaptive Deep Metric Learning for Identity-Aware Facial Expression Recognition," *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),* pp. 522-531, 2017.

[36] Z. Meng, P. Liu, J. Cai, S. Han and Y. Tong, "Identity-Aware Convolutional Neural Network for Facial Expression Recognition," *2017 12th IEEE International Conference on Automatic Face & Gesture Recognition (FG 2017),* pp. 558-565, 2017.

[37] X. Zhao, X. Liang, L. Liu, T. Li, Y. Han, N. Vasconcelos and S. Yan, "Peak-Piloted Deep Network for Facial Expression Recognition," *Computer Vision – ECCV 2016. ECCV 2016. Lecture Notes in Computer Science,* vol. 9906.

[38] K. Zhang, Y. Huang, Y. Du and L. Wang, "Facial Expression Recognition Based on Deep Evolutional Spatial-Temporal Networks," *IEEE Transactions on Image Processing,,* vol. 26, pp. 193-4203, 2017.

[39] C.-M. Kuo, S.-H. Lai and M. Sarkis, "A Compact Deep Learning Model for Robust Facial Expression Recognition," *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW),,* pp. 2202-2208, 2018.

[40] Z. Yu, Q. Liu and G. Liu, "Deeper cascaded peak-piloted network for weak expression recognition," *The Visual Computer,* 2017.

[41] W. Hua, F. Dai, L. Huang, J. Xiong and G. Gui, "HERO: Human Emotions Recognition for Realizing Intelligent Internet of Things," *IEEE Access,* pp. 24321-24332, 2019.

[42] M.-I. Georgescu, T. R. Ionescu and M. Popescu, "Local Learning With Deep and Handcrafted Features for Facial Expression Recognition," *IEEE Access,* pp. 64827-64836, 2019.

[43] T. Connie, M. Al-Shabi and P. W. Cheah, "Facial Expression Recognition Using a Hybrid CNN–SIFT Aggregator," *Multi-disciplinary Trends in Artificial Intelligence. MIWAI 2017. Lecture Notes in Computer Science,* vol. 10607, 2017.

[44] D. Li and G. Wen, "MRMR-based ensemble pruning for facial expression recognition," *Multimedia Tools and Applications,* vol. 77, pp. 1-22, 2017.

[45] B.-K. Kim, J. Roh, S.-Y. Dong and S.-Y. Lee, "Hierarchical committee of deep convolutional neural networks for robust facial expression recognition," *Journal on Multimodal User Interfaces,* vol. 10, 2016.

[46] L. Xu, M. Fei, W. Zhou and A. Yang, "Face Expression Recognition Based on Convolutional Neural," *2018 Australian & New Zealand Control Conference (ANZCC),* pp. 115-118, 2018.

[47] S. Minaee and A. Abdolrashidi, "Deep-Emotion: Facial Expression Recognition Using Attentional Convolutional Network," *arXiv preprint arXiv:1902.01019,* 2019.

[48] D. Cires¸an, U. Meier and J. Schmidhuber, "Multi-column Deep Neural Networks for Image Classification," *Proceedings / CVPR, IEEE Computer Society Conference on Computer Vision and Pattern Recognition. IEEE Computer Society Conference on Computer Vision and Pattern Recognition,* 2012.