Journal of Theoretical and Applied Information Technology
15th March 2023. Vol.101. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

IMPROVING LOW-LIGHT FACE EXPRESSION RECOGNITION USING COMBINATION OF MIRNET AND RESNET50 MODELS

RHEIVANT BOSCO THEOFFILUS[1], GEDE PUTRA KUSUMA[2]

Computer Science Department, BINUS Graduate Program - Master of Computer Science,
Bina Nusantara University, Jakarta, Indonesia, 11480

E-mail: [1]rheivant.theoffilus@binus.ac.id, [2]inegara@binus.edu

ABSTRACT

Emotion is the part of humans expressing their feeling about something. There are seven types of emotions in humans which are anger, disgust, fear, happiness, neutral, sad, and surprise. These human expressions can be examined and exploited in research fields, say, psychologists. Numerous areas, including the economic and health sectors, could benefit from this research. For instance, recognizing user expressions while a player is playing a video game in dark conditions. Other applications of the health industry include spotting furious driving expressions in cars at night. However, this topic still has certain shortcomings, one of which is illumination, which is unfortunate. Therefore, in this work, we propose a combination of low-light image enhancement and face expression recognition (FER) for recognizing expression in low-light conditions. MIRNet, RetinexNet, Retinex, and AGC are the four models that were used in this study for image enhancement. While the FER model consists of two models, ResNet50 and Inception ResNet V2. The result of the best combination of FER and image enhancement is MIRNet scratch and ResNet50 with 67.6% accuracy in low-light conditions. In our experiment, this combination of FER and image enhancement has the best accuracy.

Keywords: *Face Expression Recognition, Low Light Image, Image Enhancement, ResNet50 Model, Deep Learning.*

1. INTRODUCTION

Emotions are how human expresses what they are going through. A human can express their emotions through facial expressions, tone of voice, and body movements. Of all the ways that the human being expresses his emotions, facial expressions are the easiest way to know how humans feel by communicating. With a look on their face alone, humans can express their emotions, whether it is happy, angry, or disappointed.

Lately, facial expressions recognition is a popular topic in the fields of computer vision, image processing, machine learning, and cognitive science. Unfortunately, this topic still has some limitations, which is illumination. If we input a dataset into a machine with an image that has very low light, such as the DARK FACE dataset, the image will greatly impact the accuracy results [1] [2].

According to psychology scientists, these expressions in humans can be analyzed and used as research areas. This research can help in many ways, such as in the health sector and business. For example, detecting user expressions while playing video games [3] in dark conditions. Another example of the health sector is detecting driver expressions at night such as driving angry.

Detecting expressions in a low-light environment will be a challenge for Face Expression Recognition (FER) model. However, these limitations can be overcome by using an image enhancement technique. Image enhancement techniques can be used to set up the image's brightness. There are many image enhancement techniques & models such as the CLAHE technique [4], MirNet [5], *etc*. That technique & model is one of the solutions to increase the light in the image. But those techniques still need to be trained in various low-light environments. For example, the CLAHE technique can't set up the brightness if the image has very low lighting. In this case, we will train it

using our data so it can adjust the brightness in various lighting conditions.

Thereof the experiment used synthetic data. The data is generated from an AI generative model using the FER2013 dataset [6] that was darkened. The results of the darkened images will be used as training data.

The image enhancement will be used as image pre-processing in this research. It will help the FER model to increase its accuracy and could recognize human expression under a low-light environment. Recent studies in FER Such as Ensemble ResMaskingNet with six other CNNs [7] which is the state-of-the-art, LHC-Net [8] uses the local multi-head channel to recognize the expression, and lastly facial expression recognition with deep learning [9] model for FER. This means that even if studies have suggested a FER model, no researcher is trying to overcome the low light image problem.

Therefore, we proposed a solution for facial expression recognition with very low lighting. The research method is using a combination of two models, which are low-light image enhancement and face expression recognition. In the low-light image enhancement model, this study used four models that applied to the face expression recognition model. The 4 models are Retinex [10], Retinex-Net [11], Adaptive Gamma Correction (AGC) [12], and MirNet [5]. As for the face expression recognition model, Inception ResNet V2 [13] and lastly facial expression recognition with deep learning [9] are used. The combination of these two models is evaluated using the benchmark FER2013 dataset and measured by using PSNR, and SSIM for image enhancement. While the FER model, accuracy is used.

These two methods will be merged and used to solve the problem of facial expression recognition, which is insufficient lighting or illumination and contribute to the drawback of FER.

The contribution of this paper is to solve the following the research problems:

- Will the use of image enhancement as pre-processing improve the performance of FER/facial expression recognition?
- Which combinations of image enhancement and facial expression recognition that result in the best performance?
- Did it solve FER (Face Expression Recognition) problem in low light environment?

2. RELATED WORKS

There are two sections of related works which are, Face Expression Recognition (FER) and Image Enhancement. These related works aim to analyze the best performance model for this research.

2.1 Image Enhancement

In 2022, Fan et al. proposed a solution for low-light image enhancement using Half Wavelet Attention on M-Net+ as a method to improve image quality. In their research, Fan et al. used 4 layers, where each layer has a Half Wavelet Attention Block (HWAB) which is used to extract semantic information. In addition, this study also proposes M-Net+ which is used as an enhancement network backbone in the architecture. [14]. This experiment uses two datasets as testing, namely LOL and MIT-Adobe FiveK. For testing, this method uses only one dataset, namely the LOL dataset. As a result of this experiment, the proposed method achieved 24.44 on PSNR and 0.914 for SSIM.

Before Fan et al. proposed a solution, Wang Y. *et al* found a solution for increasing the brightness in dark images using the LLFlow method [15]. The datasets used in this experiment are LOL dataset as trains and testing, and VE LOL dataset as an evaluation dataset. This method is included in the state-of-the-art on the LOL dataset benchmark with a PSNR of 25.19 and an SSIM of 0.93.

Zhang et al. developed a method to improve image quality in dark conditions. The method used is the Prior Equalization Histogram (HEP). Two stages occur in the HEP model, where the first is the Light Up Model stage, and the second is the Noise Disentangle Module stage. Light module where the decomposition of the image occurs into map illumination and map reflectance. While the Noise Disentangle Module is a stage where the noise in the image will be reduced [16]. The dataset used is the LOL Dataset. In the evaluation results, this method received a PSNR of 20.23.

A novel method proposed by Guo et al. becomes state-of-the art on low-light image

enhancement. They use Lightweight Deep Network (LDN) and DCE-Net as proposed methods [17]. Lightweight Deep Network (LDR-Net) is a module developed from the Retinex method. The way this LDR-Networks is to estimate an illumination from the map using maximum RGB color channels which then uses Retinex theory to get reflections from the image. Then, this LDR-Net will work to refine estimates from previous reflections as well as recover lost structural and texture details. The SICE dataset [18] was used to train the DCE-Net model. This method was evaluated using NPE [19], LIME [20], MEF [21], DICM [22], and VV datasets. The evaluation results of this method are PSNR: 16.57, SSIM: 0.59, and MAE: 98.78.

Zhang *et al*. developed a solution for low-light image enhancement called KinD [23]. KinD is a method inspired by Retinex theory. The way this KinD works is to decompose the image into two components, the illumination and reflectance components. The dataset used in their model is a LOL dataset for training and testing. In the experimental results, KinD produced a PSNR of 20.8865.

Zamir et al. solve the problem of dark images. In their research, Zamir et al. entitled Learning Enriched Features for Real Image Restoration and Enhancement they used multi-scale residual blocks (MRB) as the core of MirNet. This MRB has several key elements in it, namely parallel multi-resolution convolution streams, information exchange across multi-resolution streams, attention-based aggregation of features arriving from multiple streams, dual-attention units, and residual resizing modules [24]. The datasets used to train the model are SIDD, while for testing, the datasets used are SSID and DND. In the evaluation results, this model received values of 24.14 on PSNR and 0.83 on SSIM, using the LoL dataset as its evaluation.

Retinex-Net is a model proposed by Wei et al. to improve the quality of dark images [25]. This model is a method developed from Retinex. The process of Retinex-Net is divided into three, namely decomposition, adjustment, and reconstruction. The model is trained using raw-captured images from the camera and then synthetically. During testing, the datasets used were LIME (Low Light Image Enhancement), MEF (Multi Exposure image Fusion), and DICM datasets. In his research, the results of performance measurements were not written, but we tested the performance of this model using the LOL dataset. This model gets a PSNR of 16.77 and an SSIM of 0.41909298.

In 2016, Rahman et al. proposed a method called Adaptive Gamma Correction (AGC) [12]. AGC is a method of improving the exposure of an image using calculated gamma correction. In each image, the gamma value will be calculated. This gamma value will be divided into two, namely, the value of low-light image and moderate contrast image. if the mean value of an image has an intensity smaller than 0.5 it means that the image is categorized as dark, and if the mean value of the intensity is greater than 0.5 it means that the image is categorized as bright. In his research, no dataset was used to train and test the model. Performance measurements were carried out using the LOL dataset with a PSNR of 13.60 and an SSIM of 0.25519422.

et al. researched a method of improving image quality using multi-scale Retinex with color restoration [10]. This method is done by using the variance of the histogram as a control measure, or by using the frequency of occurrence of pixels shown on the histogram. In the evaluation carried out, this method has good performance because the method used is automatic. This method improves the colors in the image automatically by using variations in the histogram and the frequency of pixels appearing in the image as a control measure. The evaluation results obtained when evaluated using the LOL dataset are SSIM of 0.471 with a PSNR of 13.40.

2.2 Face Expression Recognition

Pecoraro *et al*. in their research proposed a method of recognizing expressions on the face using the LHC-Net (Local (multi) Head Channel (self-attention) module. LHC-Net is a network where this network uses a local attention module combined with the CNN framework [8]. In its evaluation, this study used the FER2013 dataset as training and testing for the module. The results of the evaluation obtained are quite good, the accuracy obtained by this module is 74.42% in the FER2013 benchmark [6].

Khaireddin Y & Chen Z conducted a study in the field of Face Expression Recognition using the VGGNet model. The architecture used in the model uses two layers, which are a fully connected layer and a convolutional layer. There

are four stages for convolutional layers and 3 stages for fully connected layers. Each stage on the convolutional layer has two convolutional blocks and a max-pooling layer. In the convolution block there is a convolutional layer, ReLU activation, and a batch of normalization layers. Normalization layers are used to speed up the learning process, reduce internal covariance shifts, and prevent gradient loss contained in the image. So convolutional stages are responsible for features of extraction, dimensionality reduction, and nonlinearity. The first two fully connected layers are used as ReLU activation. The third layer on the fully connected layer is used as a classification. On a fully connected architecture, it is trained to classify inputs.[26]. The dataset used in this experiment is FER2013 for testing and training. In the experimental results, the accuracy result achieved by VGGNet was 73.28%.

Revina *et al*. wrote a survey study using preprocessing, feature extraction, and classification techniques in face expression recognition [27]. These techniques are the initial techniques in making face expression recognition with different methods. In addition, this paper uses the Karolinska Directed Emotional Faces Dataset as testing and training. For example, Gabor function and SVM methods achieved 99% accuracy in the Karolinska Directed Emotional Faces Dataset.

A challenge was proposed by Goodfellow *et al*. in 2013. The challenges in this study contain 3 machine learning contests, namely black box learning challenges, face expression recognition challenges, and multimodal learning challenges [6]. In 2020 Li S & Deng W took part in this challenge and managed to occupy the highest accuracy [28]. The proposed method uses network segmentation to enhance the features on the map, where this segment allows the network to focus on relevant information to make correct decisions. In their experiments, they combined deep residual networks and architectures such as U-net to produce a residual masking network. This method is trained using the FER2013 and Image-Net datasets, then tested using FER2013 and VEMO private datasets. The accuracy was 76.82%.

The facial recognition method using a CNN-based Attention Mechanism [29], has been published by Li J. *et al*. There are four parts to the proposed network which are, the feature extraction module, attention module, reconstruction module, and classification module. In this study, CNN was used as the backbone of the architecture. The method in this study used the datasets of CK+, JAFFE, FER2013, Oulu-CASIA, and self-collected Nanchang University Facial Expression (NCUFE) as training and testing. In the evaluation results, the accuracy obtained in this study was the FER2013 dataset with an accuracy of 75.82%, the CK+ dataset with an accuracy of 98.9%, JAFFE with an accuracy of 98.52%, Oulu-Casia with an accuracy of 94.63%, and NCUFE with an accuracy of 94.33%.

Vo TH. *et al*. developed a method to recognize facial expressions using Pyramid Super Resolution (PSR network. There are 6 blocks contained in the PSR architecture, including spatial trans-former network (STN), scaling, low-level feature extractor, high-level feature extractor, fully connected, and final concatenation block. STN is an affine transformation simulator for 2D images, where it is used as a face alignment. Scaling blocks are the main block or basic idea in this approach. After block scaling, there are several internal outputs, where this output is an input image that has a different scale and size. Low-level and high-level feature extractors are used to reduce the number of features in a data set by creating new features from existing ones (and then discarding the original features). In the fully connected layer block, there is a combination of a fully connected layer and a dropout layer. The last block is the output of engineering incorporation in the architecture described earlier [30]. In this experiment, FER+, and AffectNet datasets are used as trains, tests, and validation. Meanwhile, RAF-DB is only used for testing and training. The experimental results showed that this method achieved an accuracy of 88.98% in WA (Weighted Accuracy), and 80.78% in UA (Unweighted Accuracy) using RAF-DB. For the FER2013 dataset, the init method gets an accuracy of 89.75%. While the AffectNet dataset, for 8 classes, this method gets an accuracy of 60.68%, while for 7 classes the accuracy reaches 63.77%.

In 2019, Deng J et al. in their research proposed a method of detecting faces using RetinaFace [31]. RetinaFace is a robust single-stage face detector. In its evaluation, this study used the WIDERFACE dataset [32] which

contained 32,203 images. The dataset is then divided into 40% for training, 10% for validation, and 50% for testing. The evaluation results obtained are very good, whereas the AP results obtained are around 0.914%.

A method for the recognition of expression on human faces using De-expression Residue Learning has been created by Yang H. et al. In the approach carried out, there are two learning processes, namely: the first is to study neutral faces generation using cGANs, and the second is to study the process of the intermediate layer on the generator [33]. This study used BU-4DFE, BP4Dspontaneous datasets as training, CK+, Oulu-CASIA, MMI, BU3DFE, and BP4D+ as validation, and BP4D+ as testing. The evaluation experiment is divided into 2 experiments. The first experiment is to measure accuracy using the BP4D dataset as training and BP4D + as testing. For the second experiment, the model uses the BP4D+ dataset as training and testing. The evaluation results produced in the first experiment had 74.41% accuracy. While the second experiment has an accuracy of 81.39%.

Yang B. et al. conducted a study, where this study aims to enable machines to detect facial expressions using the Weighted Mixture Deep Neural Network (WMDNN) as a method to extract features that are effective in performing facial recognition automatically [34]. The evaluation was carried out using the VGG16 Network model into the method as an image extraction that has been processed by the Weighted Mixture Deep Neural Network (WMDNN) using the ImageNet Database as training. As for testing, they used the CK+ dataset, JAFFEE and OULU Casia. The accuracy obtained in each dataset is CK+ Dataset: 97.02%, JAFFE Dataset:92.21%, and OULU Casia: 92.89%.

Pramerdorfer and Kampel proposed a solution to detect expressions on faces using Convolutional Neural Networks. In their research, they combined the CNN architects they created with several architectures that fall into the state-of-the-art category [35]. The evaluation carried out in this study was to try to overcome the bottleneck that was happening by using CNN without the use of additional training data or requiring facial registration. The dataset used for training and testing is FER2013. The evaluation results obtained were in accordance with expectations, where this method can achieve an accuracy of 75.2%.

Zhao and Zhang in 2016 conducted a review and survey on the design experiment using several methods, one of which was Geometric feature-based, optical flow, and sparse representation-based classification (SRC). The feature-based geometric method is a method used to extract features in static images. For optical flow, it is used as predicting the brightness of each pixel that moves across the layer over time. This is usually applied to a series of images that have a small-time step, for example, video frames. The SRC method is used for the classification of images [36]. In the review, the evaluation was carried out using JAFFE Database with as many as 213 images and Cohn-Kane Database with as many as 320 images for training and testing. The results of the accuracy performance obtained using the SRC method are JAFFE Database by 84.76% and Cohn-Kane Database by 97.14%.

2.3 Summary of Related Works

After reviewing some of the previous work, Face Expression Recognition (FER) is still in the stage of use in normal light, or normal exposure to an image. No one has researched low illumination cases, and the FER2013 dataset has also not been made for low-light images. The best method today is FER with an ensemble model which is a combination of CNN and ResMaskingNet [7]. As for the single model research, it is LHC-Net with an accuracy of 74.42% and Residual Masking Network with an accuracy of 74.14%.

In conclusion, the problem of illumination hasn't been researched yet in the FER field. Therefore, this research focus on a low-light environment for FER. For the model to recognize facial expressions in dark conditions, We used PIL to generate the FER2013 dataset by lowering the factor of the image to darken the image to train the model. Low-light image enhancement is used as a pre-processing image. The image enhancement models are, Retinex [10], Retinex-Net [11], Adaptive Gamma Correction (AGC) [12], and MirNet [5]. As for the Facial Expression Recognition (FER) model, the experiment is using two single models for FER, which are Face Expression Recognition using Deep Learning [9] and Inception ResNet V2 [13]. The ResNet50 model will be used a pretrained model that has an accuracy of 72.4%. The second model is

Inception ResNet V2, on this model We did experimenting with some layers of the architecture in ResNet V2.

3. THEORY AND METHOD

The theory and method section will be used to describe the method and model that are used in this research. We will divide it into four sections for the model and method of image enhancement, and two sections for the FER (Facial Expression Recognition) model.

3.1 Retinex

Retinex is a theory proposed by Edwin Land [37]. This theory is a way of looking at humans seeing objects from the Retina and Cortex so it is referred to as Retinex. This retinex was created to identify the spatial image processing responsible for the constancy of colors in the image. This theory was then developed by the researchers to be able to regulate the lack of lighting in an image. The researchers did this by separating the reflection from the image that has a certain illumination.

The retinex that will be used in this study is the automatic method, where this method will adjust the lighting automatically or does not need to set the alpha value (Parthasarathy & Sankaran, 2012). This method is done by using the variance of the histogram as a control measure, or by using the frequency of occurrence of pixels shown on the histogram.

3.2 Deep Retinex Decomposition (RetinexNet)

Retinex-Net [11] is a model developed from Retinex. There are three processes in the Retinex-Net architecture, namely decomposition, adjustment, and reconstruction.

At the initial stage, the image will be decomposed using Decom-Net to break down the image into reflectance and illumination. After the image is decomposed, the image will enter the adjustment process. In the adjustment, the process image will be decomposed into illumination and sent to the encoder process, called Enhance-Net to brighten the dark lighting. Multi-scale circuits are also used to adjust the illumination of the image. Meanwhile, in the image, the reflectance can enter into a process where the noise in this image will be reduced. In the final step, these two decomposed images will be reconstructed and adjusted to get the maximum image.

3.3 MIRNet

MirNet is a model based on recursive residual design. A new architecture with Convolutional Neural Network (CNN) as a based method to improve image quality to a high-quality version by using multi-scale residual block (MRB) as the core of MirNet.

MRB in the architecture will be used to maintain a high-resolution representation, where there are five key elements in it: (a) parallel multi-resolution convolution streams to extract images from smooth to coarse and also convert images from rough to smooth semantically and precisely (b) multi-resolution cross-stream information exchange, (c) attention aggregation based on the features arriving from multi-resolution cross-stream information, (d) dual-attention units to capture contextual information in spatial and channel dimensions, and (e) residual resizing modules to perform downsampling operations (typically used to reduce storage and/or transmission requirements on images) and upsampling (increasing spatial resolution while preserving the 2D representation of an image). In addition to the MRB there is an RRG (Recursive Residual Group) on the architecture. This RRG is used to accommodate multi-scale residual blocks. [5].

3.4 Adaptive Gamma Correction

Adaptive Gamma Correction (AGC) is a method of improving the exposure of an image using calculated gamma correction. In each image, the gamma value will be calculated. This gamma value will be divided into two, the value of low-light image and moderate contrast image. If the mean value of an image has an intensity smaller than 0.5 it means that the image is categorized as dark, and if the mean value of the intensity is greater than 0.5 it means that the image is categorized as bright.

There are several steps taken in the AGC method: color transformation, image classification, and then intensity transformation. The color transformation is an image transformation carried out through three color channels contained in the image such as RGB (red-green-blue), Lab, HSV (Hue, Saturation, Value), and YUV. For the AGC model, use the color values on the HSV to separate the colors and brightness in the image into Hue, Saturation, and Value. This is because HSV also has the capacity

to represent colors for human perception and the ability to completely separate color information from brightness information in the image. Image classification is useful for classifying images. This is necessary because each image has its characteristics, and the image enhancement must be based on the characteristics of the image. Each image has a different intensity value, therefore, this classification is divided into two sub-classes, that is bright and dark. Intensity transformation is a method proposed by Rahman et al. This method contains two parameters to control the shape from the transformation curve [12].

3.5 Resnet50

The ResNet50 that was proposed by Khanzada, A., et al. achieved a state-of-the-art for FER2013 benchmark with 73.2% accuracy. In ResNet, the baseline model is initially created from scratch. Four 3x3x32 same-padding ReLU filters, two 2x2 MaxPool layers, an FC layer, and a softmax layer are used to construct it using a vanilla CNN. The batchnorm are also added about 50% droupout layers in order to handle high variation and increase their accuracy from 53.0% to 64.0%.

The ResNet50 has 50 layers in it. It has 175 layers in Keras, where it is specified. The ResNet50's original output layers is replaced by two FC layers of sizes 4,096 and 1,024 respectively and a layer of softmax output with seven emotion classes. The first 170 layers in ResNet is froze, and kept the rest of the network trainable. They also used SGD optimizer with a 0.01 of learning rate and 32 batch size. After training with 122 epochs this model achieves an accuracy of 73.2% accuracy on the test set.

3.6 Inception ResNet V2

Inception ResNet V2 architecture is taken from Keras. This architecture is usually used to solve complex problems and can improve the performance and accuracy of a model. By adding additional layers, Inception ResNetV2 can help models to learn more complex features. For example, in terms of Image Recognition [38].

There are a few steps that are taken in the process of Inception ResNet V2 architecture. After receiving the input, the image will enter a stem block that serves to learn local features such as edges in the image. After that, there is the Inception ResNet block. This block serves to upscale the dimensions on the filter to compensate for the dimensional reduction caused by the previous block as well as being used to learn and identify textures. Furthermore, there is a reduction block that serves to reduce the number of input variables in the data. Then there is an average pooling block that serves to calculate the average for each patch in the map feature. In the next block, there is a dropout block to prevent overfitting. In the last position, there is a softmax block that serves to normalize output on neural networks for adjusting their values between zero and one [13].

4. RESEARCH METHODOLOGY

4.1 Dataset

This research is using three datasets, which are the LOL dataset, the FER2013 dataset, and the FER2013 synthetic dataset. LOL dataset and FER2013 synthetic dataset are used to help training on image enhancement. While FER2013 dataset and FER2013 synthetic dataset are also used for training and evaluating the FER model.

4.1.1 LOL dataset

LOL dataset is a dataset commonly used by researchers to conduct low-light image enhancement research. This LOL dataset contains 500 pairs of images, namely low-light and normal-light. The pairs of images have been separated into two parts, 485 pair images for training and 15 pair images for testing. In the low-light image, there is noise in the image. Such noise is generated during the shooting process. Most of the images contained in the dataset are indoor images. The resolution of this dataset is the same size, which is 400 x 600. The images are shown in Figure 2.

Journal of Theoretical and Applied Information Technology
15th March 2023. Vol.101. No 5

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Figure 1: Pair Images Of LOL Dataset

Figure 2: Image Above Is The Content Of The FER2013 Dataset, Where The Labels (a) Are The Labels Angry, (b) Disgust, (c) fear, (d) Happy, (e) Neutral, (f) Sad, and (g) Surprise.

4.1.2 FER2013

The FER2013 dataset is a benchmark dataset for face expression recognition created in 2013. In the FER2013 dataset, there are about 30,000 images of faces that have different expressions. The images in the FER2013 dataset have a resolution of 48 x 48. On each image, there is its label. This label is divided into 7 types, namely 0 = Angry, 1 = Disgust, 2 = Fear, 3 = Happy, 4 = Sad, 5 = Surprise, and 6 = Neutral. However, for the disgusting label, there are only 600 images/samples, while for other expression labels, there are 5000 images/samples each. An example of an image from the FER2013 dataset can be seen in Figure 3.

4.1.3 FER2013 synthetic

The FER2013 dataset does not contain dark images, so in this study, we did a synthetic to the dataset, which is to darken the data using PIL (Python Imaging Library). PIL has a class named ImageEnhance. This class can darken the image by reducing the factor value by less than one. The more the value is down, the darker the image will be. Conversely, if the value of the factor is more than one then the brighter the image. After the data is synthesized, we use the data for training and testing. In addition to training for facial expression recognition models, we also use the FER2013 dataset as training data for low-light image enhancement models. It can be seen in

Figure 4, where the images are the result of synthetics performed using PIL enhancement.

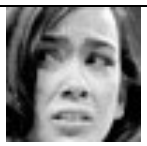
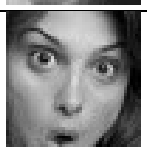
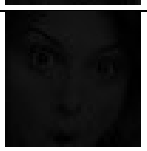
Label	Original Images	Brightness 10%
Angry		
Disgust		
Fear		
Happy		
Neutral		
Sad		
Surprise		

Figure 3: Comparison Between Original Images And Synthetic Images

4.2 Proposed Method

The proposed method of this study has two parts, namely image enhancement as an image pre-processing and facial expression recognition (FER). The proposed method of FER and image enhancement are merged in this research. Here is Table 1 which contains an overview of the combination of models and methods of image enhancement and facial expression recognition.

Table 1: Combination Of Image Enhancement And FER Model

No	Combination	
	Image Enhancement Method	Model FER
1	Retinex	ResNet50
2	RetinexNet (Scratch)	ResNet50
3	Retinex-Net (Pre-trained)	ResNet50
4	MIRNet (Scratch)	ResNet50
5	MIRNet (Pre-trained)	ResNet50
6	Adaptive Gamma Correction	ResNet50
7	Retinex	Inception ResNet V2
8	RetinexNet (Scratch)	Inception ResNet V2
9	Retinex-Net (Pre-trained)	Inception ResNet V2
10	MirNetNet (Pre-trained)	Inception ResNet V2
11	MIRNet (Scratch)	Inception ResNet V2
12	Adaptive Gamma Correction	Inception ResNet V2

Image pre-processing in this study is processing in improving image quality and lighting in an image or input. The models and methods used in this study are Retinex-Net [11] and MirNet [5]. This model was trained using the LOw-Light dataset (LOL). In the LOL dataset, there are 485 pairs of images for training and 15 pairs of images for testing. However, in this study, we changed the training, validation, and testing set from the LOL dataset and used the synthesized FER2013 dataset for the extra testing dataset. The image size/resolution of LOL and FER2013 is slightly different where LOL has a resolution of 400x600 and FER2013 has a resolution of 48x48.

After these low-light image enhancement models are trained, these models are implemented or combined into the face expression recognition module. There are two face expression recognition models that we will use, namely, ResNet50 [9] and Inception ResNet V2 [13]. The datasets that will be used for training, validation, and testing are FER2013 under normal conditions

and FER2013 in dark/synthetic conditions. The FER2013 dataset has been divided into training and testing, this division has been carried out when the dataset is downloaded. However, for this study, we changed the structure of the dataset into training, validation, and testing.

4.3 Evaluation Process

The evaluation process is divided into two stages, which are the image enhancement stage, and the face expression recognition stage.

4.4.1 Image enhancement

The image enhancement is divided into three stages to evaluate this study, which are training, validation, and testing. This training stage was carried out to train several models that can be trained used in this study, namely Retinex-Net [11], and MirNet [5]. These models will be divided into two types which are training from scratch and training using a pre-trained model. The pre-trained and scratch model is also tuned using grid-search. The hyperparameter tuning will consist of 4 tunes which are using SGD and Adam optimizer and a learning rate of 0.001 to 0.0001. The validation stage is performed to validate the model unbiasedly during hyperparameter tuning. For the testing stage, we do it to be able to measure the performance of the model and method used. For the dataset division, we made changes from the previous data, namely, set training, validation, and testing so that training images used as many as 400 pairs of images for training and 100 images for validation of the LOL dataset. After we did train on the low light image enhancement model, then the next thing we did was a test. This test will use the synthesized FER2013 dataset with a total of 7,178 images and a LOL dataset of 15 pairs of images. These images have been shared from the source.

4.4.2 Face expression recognition

Next is the evaluation of the face expression recognition (FER) side. Just like the previous stage, at this stage, three stages will be divided into training, validating, and testing. Initially, we trained the two models that will be used, which are ResNet50 [9] and Inception ResNet V2 [13]. The model will be using a pre-train model from the author. The model is also tuned with hyperparameter tuning. The ResNet50 model will consist of four hyperparameter tuning, which are using Adam optimizer and change the learning rate range from 0.001 to 0.00001. And also using SGD optimizer and changing the learning rate range from 0.01 to 0.0001. It is applied to Inception ResNet V2 too, which are using SGD and Adam optimizer with a learning rate range from 0.001 to 0.0001. The validation in this phase is carried out to perform hyper-parameter tuning by using the grid-search method to find the optimizer and learning rate of the model. The dataset that we used first to train these models is the FER2013 dataset. The dataset that we use is divided into three that is, training, validating, and testing. for training data, we used about 21,351 images. For validation, the data is taken from the rest of the training images and the number will be equal to testing, which is 7,178 images, and testing has not changed so that the data for testing is 7,178 images, but what will be used for testing is the FER2013 dataset that has been synthesized. We did the testing by trying to combine the results of the images that have been enhanced using the low-light image enhancement models that we have trained before. The selection of the best model and method will be done by comparing the performance of the combinations carried out.

5. RESULT & DISCUSSION

5.1 Image Enhancement Result

After the experiment on image enhancement, we evaluate it using PSNR and SSIM to know which model is superior. As we can see in Table 2 the MIRNet (Scratch) has the best SSIM & PSNR. MIRNet can achieve 30.03 PSNR and 0.938 SSIM. The best hyperparameter tuning for MIRNet is the SGD optimizer with a learning rate of 0.001. We can also see the qualitative of the image in Figure 5.

Table 2: Evaluation Result Of Image Enhancement Model Quantitative

Methods	PSNR (dB)	SSIM
MIRNet (Scratch)	30.03	0.938
MIRNet (Pretrained)	29.91	0.935
RetinexNet (Scratch)	16.75	0.729
RetinexNet (Pretrained)	16.75	0.729
AGC	13.70	0.594
Retinex	12.50	0.735
Darkened Images	6.321	0.062

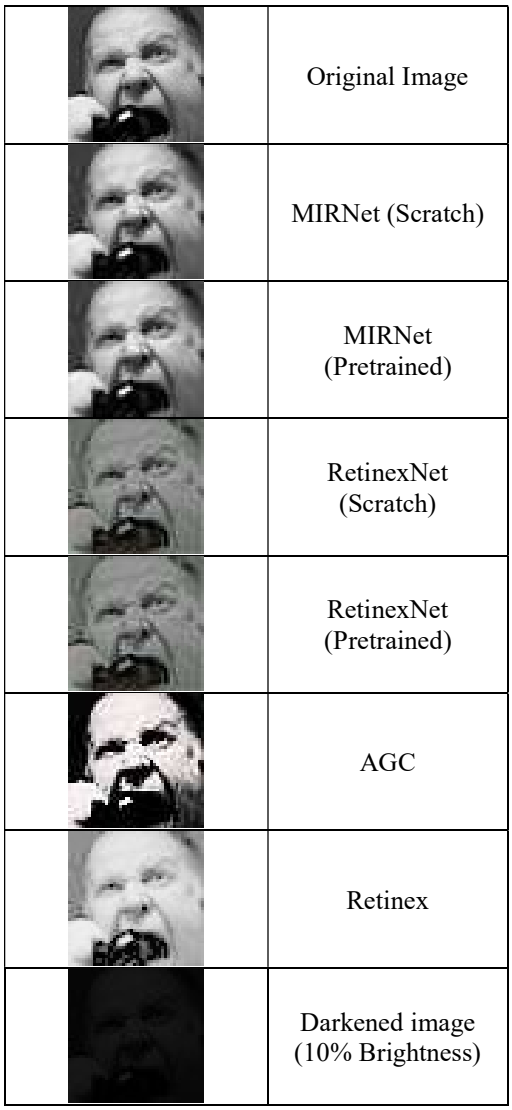


Figure 4: Evaluation Result Of Image Enhancement Model Qualitative

5.2 FER Result

The experiment we do in the FER section has 16 methods from the combination of the image enhancement method and the FER model. Method 1 and 8 is a predicted results of Inception ResNet V2 using original images and synthetic images. Method 9 and method 16 also a predicted results of ResNet50 using original images and synthetic images.

5.2.1 Training & Validation Results for FER

Table 3 is sorted in descending according to validation accuracy. As we can see in Table 4, Inception ResNet V2 (RV2) has the best training accuracy and validation accuracy on the original and darkened dataset.

Table 3: Training Result Of FER On original Dataset And Darkened Images

Method	Train Acc (%)	Train Loss	Val Acc (%)	Val Loss
RV2 x Original	77.9	0.623	77.9	0.612
R50 x Original	76.2	0.783	71.6	0.867
RV2 x Dark	70.0	0.846	67.9	0.906
R50 x Dark	67.1	0.873	56.1	1.178

From Table 4, we shorten the method's name so it can fit in the table and sort it descending according to validation accuracy. The FER model is ResNet50 (R50) and Inception ResNet V2 (RV2). While the image enhancement methods are MIRNet scratch (MNet Scr), MIRNet Pretrained (MNet Pre), RetinexNet scratch (RNet Scr), RetinexNet Pretrained (RNet Pre), Retinex, and AGC.

In Table 4, we can see that the best train accuracy from the combination model is a combination model of ResNet V2 and MIRNet pre-trained with 70.4% accuracy & 0.877 loss.

Table 4: Training & Validation Result Using Combination Of FER & Image Enhancement

Method	Train Acc (%)	Train Loss	Val Acc (%)	Val Loss
RV2 x MNet Pre	**70.4**	**0.833**	**68.8**	**0.877**
RV2 x MNet Scr	70.6	0.820	68.6	0.866
R50 x MNet Scr	68.5	0.946	64.0	1.013
R50 x MNet Pre	67.9	0.953	63.9	1.030
R50 x AGC	51.1	1.300	60.3	1.108
R50 x RNet Pre	60.4	1.109	58.7	1.163
R50 x RNet Scr	60.4	1.109	58.6	1.159
RV2 x Retinex	58.0	1.142	58.0	1.140
RV2 x RNet Scr	58.2	1.168	57.7	1.192
RV2 x RNet Pre	57.9	1.172	57.0	1.210

R50 x Retinex	53.1	1.268	51.4	1.296
RV2 x AGC	44.9	1.862	44.6	1.843

5.2.2 Testing result for FER

Table 5 shows that ResNet50 yielded the best results for testing in normal and low-light condition, with an accuracy of 70.7% and 64.8% respectively.

Table 5: Testing Result Of FER On Original Dataset And Darkened Images

Method	Test Acc (%)
R50 x Original	70.7
RV2 x Original	68.9
R50 x Dark	64.8
RV2 x Dark	58.0

Eventhough in Table 4, the best combination with the highest train and validation accuracy is MIRNet pre-trained (MNet Pre) and Inception ResNet V2 (RV2). However, from Table 6, the best accuracy of the method combination of FER and image enhancement is held by MIRNet scratch (MNet Scr) and ResNet50 (R50) with an accuracy of 67.6 %.

Table 6: Testing Result Using Combination Of FER & Image Enhancement

Method	Test Acc (%)
R50 x MNet Scr	**67.6**
R50 x MNet Pre	66.7
R50 x Retinex	66.0
RV2 x MNet Scr	62.5
RV2 x MNet Pre	62.0
R50 x RNet Scr	62.0
R50 x RNet Pre	62.0
RV2 x Retinex	59.0
R50 x AGC	57.1
RV2 x RNet Scr	55.6
RV2 x RNet Pre	55.6
RV2 x AGC	41.0

5.3 Discussion

We can observe from the outcomes of our studies that, the combination of FER and image enhancement can improve the accuracy of FER models. In this experiment, MIRNet scratch has the best PSNR and SSIM than the other methods. MIRNet can achieve 30.03 PSNR and 0.938 SSIM. This combination made the best result in this experiment. As we can see the ResNet50 can yield an accuracy of 67.6%. While in the low-light dataset, ResNet50 can only achieve 64.8% accuracy.

This study is the first of its kind to evaluate the use of image enhancement as pre-processing to improve performance of Facial Expression Recognition (FER) in dark environment. Besides that, this study also aiming to find the best combination method for image enhancement and FER model which is MIRNet scratch and ResNet50 in this experiment. This also answer that image enhancement can solve FER problem in recognition and classified human expression.

6. CONCLUSION & FUTURE WORKS

In this research, we tested several image enhancement methods and face expression recognition models. The method that we used in this research is ResNet50 and Inception ResNet V2 for FER models. MIRNet, RetinexNet, AGC, and Retinex are also used for image enhancement methods. After all the experiments we did in this research we can conclude that image enhancement can improve the accuracy of the FER model in low-light conditions. This is because we are using image enhancement as image pre-processing to help FER models to recognize it better. However, we found that using MIRNet pre-trained with Inception ResNet V2 achieve better accuracy in training and validation. But, in testing section, the best accuracy is yielded by MIRNet using ResNet50 scratch as image pre-processing. This means that, even best accuracy in training section doesn't mean it is the best model. Therefore, we conclude that, in this experiment, The best combination of FER and image enhancement is MIRNet scratch with ResNet50. Without using image pre-processing, ResNet50 can only achieve 64.8% in low-light conditions. While using an image enhancement, the MIRNet model with SSIM of 30.03 and SSIM of 0.938 as image pre-processing, ResNet50 can get 67.6% of accuracy. In our experiment, this combination of FER and image enhancement has the best accuracy. This also means that using image enhancement for image preprocessing can solve FER problem in low-light condition. Unfortunately, many models from the previous works we stated before did not utilize image

enhancement to elevate the accuracy. Using image enhancement as image preprocessing in FER model can elevate the accuracy about 2.8%, which is from 64.8% to 67.6% in low-light condition.

However, our experiment still had some limitation such as the dataset that we used in this experiment is a synthetic dataset. Also, this paper focused only on accuracy. In the future, it is also necessary to develop a combination model in real-time, such as creating a mobile web application with real-time recognition speeds that can be used in the actual world using the best image enhancement and FER models.

REFERENCES

[1] Wei C, Wang W, Yang W, Liu J. Deep RETINEX decomposition for low-light enhancement. arXiv.org. 2018 [cited 2022Dec16]. Available from: https://arxiv.org/abs/1808.04560

[2] Yang W, Yuan Y, Ren W, Liu J, Scheirer WJ, Wang Z, et al. Advancing Image Understanding in Poor Visibility Environments: A Collective Benchmark Study. IEEE Transactions on Image Processing. 2020;29:5737–52.

[3] Ko B. A Brief Review of Facial Emotion Recognition Based on Visual Information. Sensors [Internet]. 2018 Jan 30;18(2):401. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5856145/

[4] Rahman MF, Sthevanie F, Ramadhani KN. Face Recognition In Low Lighting Conditions Using Fisherface Method And CLAHE Techniques [Internet]. IEEE Xplore. 2020 [cited 2022 Dec 16]. p. 1–6. Available from: https://ieeexplore.ieee.org/abstract/document/9166317/

[5] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, et al. Learning Enriched Features for Real Image Restoration and Enhancement. Computer Vision – ECCV 2020. 2020;492–511.

[6] Challenges in Representation Learning: Facial Expression Recognition Challenge | Kaggle [Internet]. Kaggle.com. 2013. Available from: https://www.kaggle.com/c/challenges-in-representation-learning-facial-expression-recognition-challenge/data.

[7] Li S, Deng W. Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. 2020;1–1.

[8] Pecoraro R, Basile V, Bono V. Local Multi-Head Channel Self-Attention for Facial Expression Recognition. Information. 2022 Sep 6;13(9):419.

[9] Khanzada A, Bai C, Celepcikay FT. Facial Expression Recognition with Deep Learning. arXiv:200411823 [cs] [Internet]. 2020 Apr 7; Available from: https://arxiv.org/abs/2004.11823

[10] Parthasarathy S, Sankaran P. An automated multi Scale Retinex with Color Restoration for image enhancement. In: 2012 National Conference on Communications (NCC). IEEE; 2012. p. 1–5.

[11] Wei C, Wang W, Yang W, Liu J. Deep Retinex Decomposition for Low-Light Enhancement. arXiv:180804560 [cs] [Internet]. 2018 Aug 14; Available from: https://arxiv.org/abs/1808.04560

[12] Rahman S, Rahman MM, Abdullah-Al-Wadud M, Al-Quaderi GD, Shoyaib M. An adaptive gamma correction for image enhancement. EURASIP Journal on Image and Video Processing. 2016 Oct 18;2016(1).

[13] Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning [Internet]. www.aaai.org. 2017. Available from: https://www.aaai.org/ocs/index.php/AAAI/AAAI17/paper/viewPaper/14806

[14] Fan C-M, Liu T-J, Liu K-H. Half Wavelet Attention on M-Net+ for Low-Light Image Enhancement. arXiv:220301296 [cs, eess] [Internet]. 2022 Mar 2 [cited 2022 Dec 16]; Available from: https://arxiv.org/abs/2203.01296

[15] Wang Y, Wan R, Yang W, Li H, Chau L-P, Kot A. Low-Light Image Enhancement with Normalizing Flow. Proceedings of the AAAI Conference on Artificial Intelligence [Internet]. 2022 Jun 28 [cited 2022 Sep 25];36(3):2604–12. Available from: https://ojs.aaai.org/index.php/AAAI/article/download/20162/19921

[16] Zhang F, Shao Y, Sun Y, Zhu K, Gao C, Sang N. Unsupervised Low-Light Image Enhancement via Histogram

Equalization Prior. arXiv:211201766 [cs, eess] [Internet]. 2021 Dec 3 [cited 2022 Dec 16]; Available from: https://arxiv.org/abs/2112.01766

[17] Guo C, Li C, Guo J, Loy CC, Hou J, Kwong S, et al. Zero-Reference Deep Curve Estimation for Low-Light Image Enhancement [Internet]. openaccess.thecvf.com. 2020 [cited 2022 Dec 16]. p. 1780–9. Available from: http://openaccess.thecvf.com/content_CVPR_2020/html/Guo_Zero-Reference_Deep_Curve_Estimation_for_Low-Light_Image_Enhancement_CVPR_2020_paper.html

[18] Cai J, Gu S, Zhang L. Learning a Deep Single Image Contrast Enhancer from Multi-Exposure Images. IEEE Transactions on Image Processing. 2018 Apr;27(4):2049–62.

[19] Wang S, Zheng J, Hu H-M, Li B. Naturalness Preserved Enhancement Algorithm for Non-Uniform Illumination Images. IEEE Transactions on Image Processing. 2013 Sep;22(9):3538–48.

[20] Guo X, Li Y, Ling H. LIME: Low-Light Image Enhancement via Illumination Map Estimation. IEEE Transactions on Image Processing. 2017 Feb;26(2):982–93.

[21] Ma K, Kai Zeng, Zhou Wang. Perceptual Quality Assessment for Multi-Exposure Image Fusion. IEEE Transactions on Image Processing [Internet]. 2015 Nov [cited 2019 Jul 24];24(11):3345–56. Available from: https://ece.uwaterloo.ca/~k29ma/papers/15_TIP_MEF.pdf

[22] Lee C, Lee C, Kim C-S. Contrast Enhancement Based on Layered Difference Representation of 2D Histograms. IEEE Transactions on Image Processing. 2013 Dec;22(12):5372–84.

[23] Zhang Y, Zhang J, Guo X. Kindling the darkness: A practical low-light image enhancer. In: Proceedings of the 27th ACM International Conference on Multimedia. New York, NY, USA: ACM; 2019.

[24] Zamir SW, Arora A, Khan S, Hayat M, Khan FS, Yang M-H, et al. Learning Enriched Features for Real Image Restoration and Enhancement. Computer Vision – ECCV 2020. 2020;492–511.

[25] Wei C, Wang W, Yang W, Liu J. Deep Retinex Decomposition for Low-Light Enhancement. arXiv:180804560 [cs] [Internet]. 2018 Aug 14; Available from: https://arxiv.org/abs/1808.04560

[26] Khaireddin Y, Chen Z. Facial Emotion Recognition: State of the Art Performance on FER2013. arXiv:210503588 [cs] [Internet]. 2021 May 8 [cited 2021 Nov 21]; Available from: https://arxiv.org/abs/2105.03588

[27] Revina IMichael, Emmanuel WRS. A Survey on Human Face Expression Recognition Techniques. Journal of King Saud University - Computer and Information Sciences [Internet]. 2018 Sep; Available from: https://www.sciencedirect.com/science/article/pii/S1319157818303379

[28] Li S, Deng W. Deep Facial Expression Recognition: A Survey. IEEE Transactions on Affective Computing. 2020;1–1.

[29] Li J, Jin K, Zhou D, Kubota N, Ju Z. Attention mechanism-based CNN for facial expression recognition. Neurocomputing. 2020 Oct;411:340–50.

[30] Vo T, Lee G, Yang H, Kim S. Pyramid With Super Resolution for In-the-Wild Facial Expression Recognition. IEEE Access [Internet]. 2020 [cited 2021 Apr 2];8:131988–2001. Available from: https://ieeexplore.ieee.org/document/9143068

[31] Deng J, Guo J, Ververas E, Kotsia I, Zafeiriou S. RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild. In: 2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2020. p. 5203–12.

[32] Yang S, Luo P, Loy CC, Tang X. WIDER FACE: A face detection benchmark. In: 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). IEEE; 2016. p. 5525–33.

[33] Yang H, Ciftci U, Yin L. Facial expression recognition by DE-expression residue learning. In: 2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition. IEEE; 2018. p. 2168–77.

[34] Yang B, Cao J, Ni R, Zhang Y. Facial Expression Recognition Using Weighted Mixture Deep Neural Network Based on

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Double-Channel Facial Images. IEEE Access. 2018;6:4630–40.

[35] Pramerdorfer C, Kampel M. Facial Expression Recognition using Convolutional Neural Networks: State of the Art. arXiv:161202903 [cs] [Internet]. 2016 Dec 8 [cited 2021 Apr 2]; Available from: https://arxiv.org/abs/1612.02903

[36] ZHAO, Xiaoming; ZHANG, Shiqing. A review on facial expression recognition: feature extraction and classification. IETE Technical Review, 2016, 33.5: 505-517.

[37] McCann, John. "Retinex theory." Encyclopedia of Color Science and Technology (2016): 1118-1125.

[38] Mujtaba H. What is Resnet or Residual Network | How Resnet Helps? [Internet]. GreatLearning Blog: Free Resources what Matters to shape your Career! 2020. Available from: https://www.mygreatlearning.com/blog/resnet/#sh1