Human Face Emotions Recognition from Thermal Images Using DenseNet

Original Scientific Paper

S. Babu Rajendra Prasad
VIT-AP University
School of CSE, Amaravathi, Vijayawada,
Andrapradesh, India.
baburajendraprasad655@gmail.com

B. Sai Chandana
VIT-AP University
School of CSE, Amaravathi, Vijayawada,
Andrapradesh, India.
saichandanas869@gmail.com

Abstract *– In the current scenario face identification and recognition is an important technique in surveillance. The face is a necessary biometric in humans. Therefore face detection plays a major job in computer vision applications. Several face recognition and emotions classification approaches have been presented throughout the last few decades of research to improve the rate of face recognition for thermal pictures. However, in real-time, lighting conditions might change due to several factors, such as the different times of capture, weather, etc. Due to variations in lighting intensity, the performance of the facial expression recognition system is not good. This paper proposed a model for human thermal face detection and expression classification. Four main steps were involved in this research. Initially, the Difference of the Gaussian (DOG) filter is utilized to crop the input thermal images and then normalize the images using the median filter in pre-processing step. Then, Efficient Net is used for extracting features such as shape, location, and occurrences from thermal face images. After that, detect human faces utilized by the YOLOv4 technique to better emotions classification. Finally, classify the emotions on faces by using the DenseNet technique into seven emotions such as happy, sad, disgust, surprise, anger, fear, and neutral. The proposed method outperforms state-of-art techniques for face recognition on thermal pictures, and classifies the expressions, according to experimentations on the RGB-D-T database. The accuracy, precision, recall, and f1-score metrics will be utilized with the database to assess the efficacy of the proposed methodology. The proposed models achieve a high classification accuracy of 95.97% on the RGB-D-T database. Furthermore, the outcomes show good precision for various face recognition tasks.*

Keywords*: Deep learning, detection, face expressions classification, feature extraction, pre-processing, thermal face recognition, and thermal image.*

1. INTRODUCTION

Recently, emotion detection has been used for a variety of purposes, including job interviews to identify whether a candidate is at ease, frightened, or confident, classrooms to assess whether students are paying attention, and supermarkets to consumer behavior with purchases. The primary concept behind emotion identification is based on features of the face, such as the shape of the mouth and eyes. To develop an emotion-detecting system, those expressions are crucial. The main problem is figuring out how to extract those expressions from high-resolution images where the face only takes up around 10% of the overall image space.

Face detection has numerous uses in the fields of information security, video surveillance, and identity authentication. The visible spectrum has received the majority of attention in face identification, this is dependent on outside factors like lighting. Measuring the light reflected by the face is necessary for visible spectrum imaging. As a result, changes in illumination can affect visual appearance significantly and impair the functioning of such devices [1]. Visible face identification is still difficult, mostly because of the many environmental changes that affect it, like poor lighting, uneven illumination, and viewing angles. In addition, hackers can recreate facial patterns to fool visible face detection systems. Since

thermal imaging records the heat radiation from the face and body temperature even in a completely dark area, it has been suggested as an answer in the latest years [2-4]. The likelihood of creating a fake face pattern is greatly diminished since the temperature of the skin on the face is directly tied to the underlying blood vessels that are specific to every unique person [5]. The performance of thermal face identification is, nonetheless, hampered by many problems, including pictures with a lot of noise, opacity to glass, low resolution, and sensitivity to temperature variation.

The last few years have seen the development of thermal IR imaging-based face recognition as a promising addition to traditional visible spectrum-based methods. Depending on their temperature and properties, several things emit a range of infrared energy [6-8]. The human body and face temperatures have a range that is very consistent and similar. As a result, the thermal signature is constant. Thermal emissivity from the facial surface is measured using IR cameras, and their images are largely stable under changes in lighting. Subsurface characteristics are included in the anatomical data that infrared technology images. The objective of thermal face detection is to recognize a person who has been photographed in the thermal spectrum by comparing them to visible spectrum face photographs that are the most similar [9]. Thermograms, also known as thermal images are created using IR radiation by the thermal infrared (IR) camera as it picks up the heat that the surveillance target emits. These images are used in surveillance systems for recognition (e.g., the ability to categorize an object into one of the classes like a vehicle, human, or animal), the ability to define surveilled objects in detail (like a woman with a hat, a bear, a man with a coat, etc.), and object detection (e.g., the ability to tell an object apart from the backdrop). Since the camera can catch a face from a set distance away, it is a non-intrusive method of identification [10-12]. Thermal face biometrics can help with the difficult process of identifying or verifying individuals only based on their thermal characteristics.

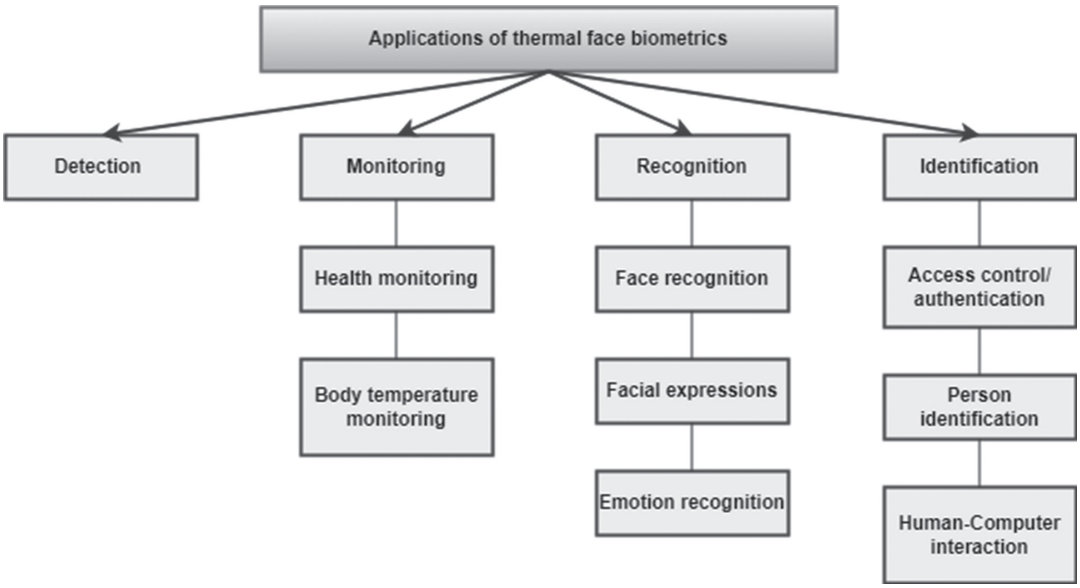

Fig. 1. Thermal biometrics applications

There are numerous applications for thermal face biometrics, including access control to protect computer systems and recognition and facilities like financial transactions, authentication for private banking, government buildings and automatic screening for terrorists at airports, use of surveillance footage, tracking body temperature in medical diagnostics, recognition of facial expression in high-end security applications, etc. Figure 1, De-identification techniques are being developed concurrently because privacy protection is a priority due to the prevalence of surveillance cameras.

A deep learning approach to thermal images is tried to solve all the existing challenges [13]. Deep learning-based techniques have been well-known in the research community as a result of their improved performance in recognizing a variety of difficulties, such as object identification, face recognition, handwritten digit recognition, speech recognition, and human action recognition. The proposed method uses the DenseNet model, which is claimed to be superior to other approaches that use the CNN, SVM, ANN, and YOLO models. The paper's contributions are,

- In the pre-processing step, cropped the input thermal face images using by Difference of the Gaussian filter and then used the median filter to normalize the lighting variations of input thermal images.
- To extract the features such as positioning, shape, and the presence of a face from thermal face images, utilizing the Efficient Net technique.
- The YOLOv4 technique is proposed to detect the face using collected features to better classification of facial emotions.

- To classify facial expressions like happy, sad, disgusted, surprised, angry, fearful, and neutral from thermal face images, we utilized the DenseNet methodology.
- This paper proposes a method for detecting several facial emotions in thermal pictures utilizing the RGB-D-T database for the dataset of human facial expression recognition.
- The proposed model performs significantly better in the identification of the item and is capable of extracting information from photos.

The remainder of the paper is ordered as follows. Section 2 presents the related works in the paper. In section 3, the problem statement is mentioned. The proposed methodology is shown in section 4. In section 5, the result section is shown. And finally, in Section 6, the conclusion is presented.

2. LITERATURE REVIEW

A literature review is presented and discussed in this paper to offer a speculative background about thermal images and human thermal face emotions recognition methodologies.

To identify human targets in aerial view thermal pictures, Akshatha et al. [14] suggested Faster R-CNN and single-shot multi-box detector (SSD) algorithms with various backbone networks. To achieve this, two common aerial thermal datasets with variously sized human objects ResNet50, Inception-v2, and MobileNet-v1 are taken into consideration. By having a mean average precision (mAP at 0.5 IoU) of 100% for the test data from the OSU thermal dataset and 55.7% for the test data from the AAU PD T datasets, respectively, the evaluation results show that the Faster R-CNN model trained with the ResNet50 network architecture outperformed in terms of detection accuracy. The suggested framework outperforms some cutting-edge approaches compared here with a recognition accuracy of 87.46%.

Bhattacharyya et al. [15] suggested the use of the effective deep learning model IRFacExNet for the detection of facial expressions in thermal pictures. According to the direction of the research, they used a DCNN structure to build this model. To extract usable information from human faces that can be utilized for the recognition of different expressions, they have utilized two structural units, the transformation unit and the residual unit, each of which has unique strengths. The suggested framework outperforms some cutting-edge approaches compared here with a recognition accuracy of 81.16%.

Nayak et al. [16] suggested a three-stage HCI system for processing multivariate time-series thermal video sequences to identify human emotion and provide diversion recommendations. The first stage consists of following the face ROIs throughout the thermal video while simultaneously detecting faces, eyes, and noses using a Faster R-CNN (region-based convolutional neural network) structure. The multivariate time series (MTS) data is created by calculating the mean intensity of ROIs. To categorize the emotional states induced by video stimulation, the smoothed MTS data are then sent to the Dynamic Time Warping (DTW) algorithm. In the third stage of HCI, the suggested framework offers pertinent recommendations from a physical and psychological distraction perspective. Both their created data set and the NVIE data set show 93.5% accuracy with the suggested Faster R-CNN architecture.

A method of thermal-visible face detection was suggested by Kamel et al. [17]. This approach, which is based on the YOLO v3 framework, offers enhanced solutions for face detection in both visual and thermal imaging, making it appropriate for a variety of applications including facial emotion recognition (FER) or liveness detection. The second is that they fully labeled a thermal face database and made it available to the scientific community in a GitHub repository. Thirdly, they have presented TVCycleGAN, a modified version of CycleGAN that enables the conversion of LWIR images into visible-like visuals. Finally, the networks synthesized-visible face images show great promise for thermal facial landmark identification. The suggested framework outperforms some cutting-edge approaches compared here with a recognition accuracy of 89.27%.

Siddiqui et al. [18] introduced a multimodel automated emotion recognition (AER) that is very accurate at discriminating between emotional expressions. The contribution comprises integrating speech with visible and infrared (IR) images to build an ensemble-based strategy for the AER. The architecture is implemented in two layers, with the first layer employing a single modality to identify emotions and the second layer fusing the modalities to categorize feelings. The classification and feature extraction processes have been carried out using convolutional neural networks (CNN). To merge the features and the decisions at various phases, a hybrid fusion strategy was used, consisting of late (decision-level) and early (feature-level) fusion. To arrive at the final determination, the output of the images classifier and the output of the CNN both of which were trained using speech samples from the RAVDESS database were combined. Similar f-score (0.87), precision (0.88), and recall (0.86) scores as well as an accuracy of 86.36% were attained.

3. PROBLEM STATEMENT

Visible cameras can be quite helpful in daytime situations, but they have significant problems with human face detection that may limit their utility. The difficulties are as follows:

- The street lights during the night may make it harder to see people.
- Visibility may be reduced by obstructions like trees, buildings, and cars.

- When there are many occlusions between humans or blurring as a result of cameras closing down gradually when they are placed on moving vehicles, it is very challenging to detect humans.
- It can be challenging to identify humans when the weather is foggy, rainy, snowy, or melting snow.

4. PROPOSED METHODOLOGY

Fig. 2 shows the proposed technique for the classification of the emotions of a person's face-based DenseNet technique. To recognize the human face in a thermal image, four steps are presented in this paper. They are pre-processing, feature extraction, detection, and classification. In a pre-processing step, cropped the input thermal face images using by Difference of Gaussian (DOG) filter and used the median filter to normalize the lighting variations of input thermal images. After pre-processing step the generated output images are ready to feature extraction. An Efficient Net approach is proposed to extract the features from input thermal face images. It extracts the features of the face, such as positioning, shape, and the presence of a specific object. Then YOLOV4 technique is used to detect the human face from extracted features. And then to classify the emotions on faces in thermal images, we used the DenseNet technique. For this experiment, we collect the images from the RGB-D-T database.

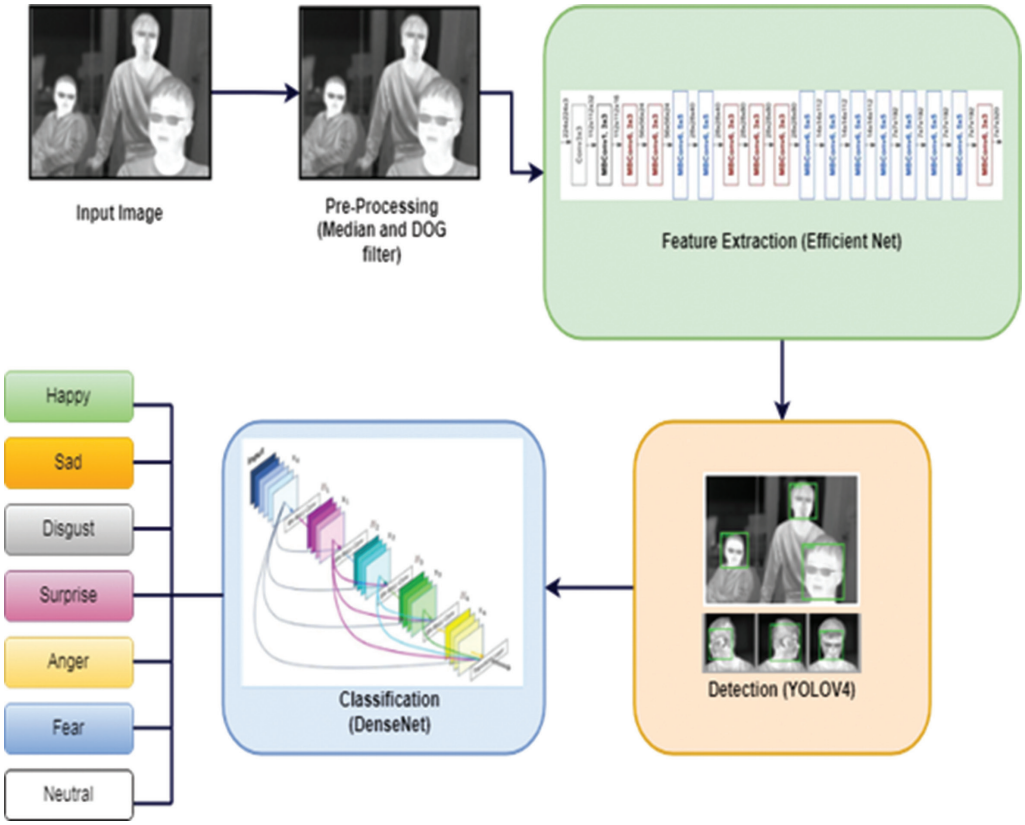

Fig. 2. The architecture of the proposed method

Pre-processing is a procedure that must be taken to extract useful information from face images. The Difference of Gaussian (DOG) filtering, a widely used method to reduce lighting variations for visible face detection, is then applied to the cropped thermal face images. An initial image and a DOG filter are convoluted to create a DOG filter image. A two-dimensional DOG filter is used for images. The one dimension is

$$G(u) = \frac{1}{\sqrt{2\pi}}\left(\frac{exp\left(-(u+ú)^2/2\sigma_1^2\right)}{\sigma_1} - \frac{exp\left(-(u+ú)^2/2\sigma_2^2\right)}{\sigma_2}\right) \quad (1)$$

And in two dimensions it is

$$G(r) = \frac{1}{2\pi}\left(\frac{exp\left(-(r+P)^2/2\sigma_1^2\right)}{\sigma_1^2} - \frac{exp\left(-(r+P)^2/2\sigma_2^2\right)}{\sigma_1^2}\right) \quad (2)$$

In addition to reducing local changes in the thermal imaging that result from the changing heat/temperature spread of the face, DOG filtering also minimizes lighting varieties in thermal facial imagery. Therefore, by lowering the local fluctuations in each technique while improving edge details, DOG filtering aids in closing the modality gap. The Gaussian filter is carefully chosen to include frequencies that are helpful for face recognition and reduce those frequencies that are negative for it. To lessen the variations and lighting in the face image the median filter can be used for normalization.

$$\hat{f}(x,y) = \underset{(s,t)\in S_{xy}}{\text{median}}\{g(s,t)\} \quad (3)$$

4.2. FEATURE EXTRACTION USING EFFICIENTNET

After pre-processing the images, we need to extract the features of the face for better prediction of the face

for classifying the emotions. The best shape information is discovered by feature extraction. Using these characteristics to categorize activities is simple with a systematic process. Features may appear differently to people and robots because they can be understood by machines. Almost all features serve to convey one aspect of an image, such as its shape, location, and occurrence of a certain face. Original images must first go through some preprocessing steps to ensure that they are suitable for feature extraction before they can begin.

After pre-processing the images, the resulting image is fed to the feature extraction method for extracting the facial feature using Efficient Net. Each of the modalities comprises a huge number of slices that form the segments using the Efficient Net technique.

The architecture of EfficientNet originates from the compound scaling method. The family of EfficientNet is built on the baseline model EfficientNet-B0 (i.e., $\varphi = 0$). The structure of EfficientNet-B0 is summarized in Table 1. An artificial neural network's performance can be improved by carefully balancing network width, resolution, and depth. Compound scaling necessitates balance and coordination between the three scaling dimensions because they are not independent. For this, a new baseline structure called EfficientNetB0 was created, which was then scaled up using compound scaling to produce the EfficientNetB0 to the EfficientNetB7 family of EffcientNets [19]. The feature extraction from the EfficientNet has been effectively combined in the proposed study. The acquired findings are comparable to those of cutting-edge networks.

Table 1. The EfficientNet [19] Framework

Level	Operator	Resolution	Channels	Layers
EfficientNetB0 architecture, the network baseline				
1	Conv1×1/Pool/FC	7×7	1,280	1
2	MBConv6, k3×3	7×7	320	1
3	MBConv6, k5×5	14×14	192	4
4	MBConv6, k5×5	14×14	112	3
5	MBConv6, k3×3	28×28	80	3
6	MBConv6, k5×5	56×56	40	2
7	MBConv6, k3×3	112×112	24	2
8	MBConv1, k3×3	112×112	16	1
9	Conv 3×3	224×224	32	1
Additional layers				
10	FC/Softmax	1	NC	1
11	FC/BN/Swish	1	128	1
12	FC/BN/Swish/Dropout	1	512	1
13	B.N./Dropout	7×7	1280	1

EfficientNet presents a new method for scaling network proportions by consistently scaling all resolution variables and structure depth/width utilizing the compound coefficient, a highly efficient agent. When compared to other CNN methods, EfficientNet provides a highly effective technique to improve accuracy while being more efficient. The EfficientNet family includes many distinct versions that are adapted to different input layers. In several dimensions, this DCNN has been built up [20]. Most CNN architectures are built up by including more layers as part of the ResNet family. Unlike previous methods, EfficientNet built up all width, depth, and resolution dimensions simultaneously. The compound scaling mechanism was introduced in the proposed EfficientNet structure to balance all of these scaling characteristics.

Convolutional neural networks belong to the EfficientNet family. In terms of layer depth, input resolution, layer width, and a combination of these criteria, EfficientNet models scale well. EfficientNet is a recent deep-learning model that aims to improve model efficiency while also improving accuracy. From B0 to B7, there are various variants. This network's fundamental building piece is MBConv, which has compression and excitation optimization added to it. Between the starting and finish of a convolutional block, these blocks provide shortcuts. To improve the depth of the feature maps, the input activation maps are enlarged using 1x1 convolutions. The thin layers are connected by shortcut links in this paradigm, whereas the broader layers are positioned between the jump links. This structure aids in the reduction of both the model's size and the overall number of transactions needed. The EfficientNetB0 structure is shown in Fig. 3.

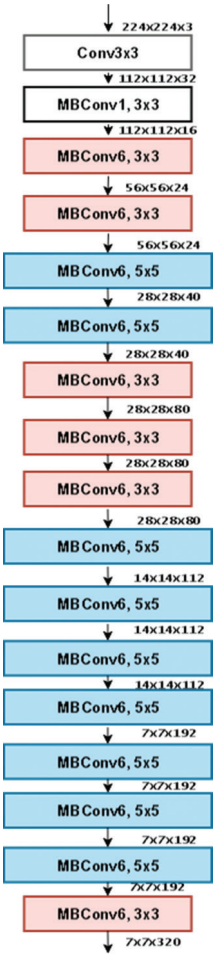

Fig. 3. EfficientNet B0 architecture [21]

Because of its higher prediction performance using a compound scaling technique across all parameters of the networks, such as depth, resolution, and breadth, EfficientNets has gotten a lot of attention [21]. To be clear, width defines the number of dimensions in any depth, the resolution is the size of the image, and the layer represents the number of levels in CNN. The theory behind compound scaling is that increasing any network dimension (width, image resolution, or depth) improves accuracy, but as the model grows larger, the accuracy gain diminishes. Compound scaling employs a compound coefficient to govern how many extra resources are useful for model scaling, and the parameters are scaled in the following fashion by the compound coefficient:

$$\begin{gathered} Resolution\ R\ \gamma\ \Phi \\ Width\ w\ \beta\Phi \\ While\ \alpha\beta 2\gamma 2 \approx 2 \\ \alpha \geq 1, \beta \geq 1, \gamma \geq 1. \end{gathered} \tag{4}$$

where the grid search determines the constants α, β, γ. The compound scaling yields the following coefficient after numerous experiments and considerable grid search:

Depth 1.20 / Width 1.10 / Resolution 1.15

In this feature extraction stage, it's extracting the features such as their shape, location, and occurrence to detect the face in thermal human face images.

4.3. YOLOV4 FOR DETECTION

YOLOV4 is proposed to detect the human face. After extracting the facial feature in the feature extraction stage, the detection methodology is ready to detect the faces in thermal images. This face detection step is for classifying the facial emotions clearly and providing better classification accuracy. As shown in Fig. 4, YOLOv4 is a single-stage detector that classifies and efficiently localizes the objects in an image in one pass. It was published in April 2020 and featured several data augmentation strategies, pre-and post-processing approaches, as well as minor model adjustments. The following is a concise description of YOLOv4.

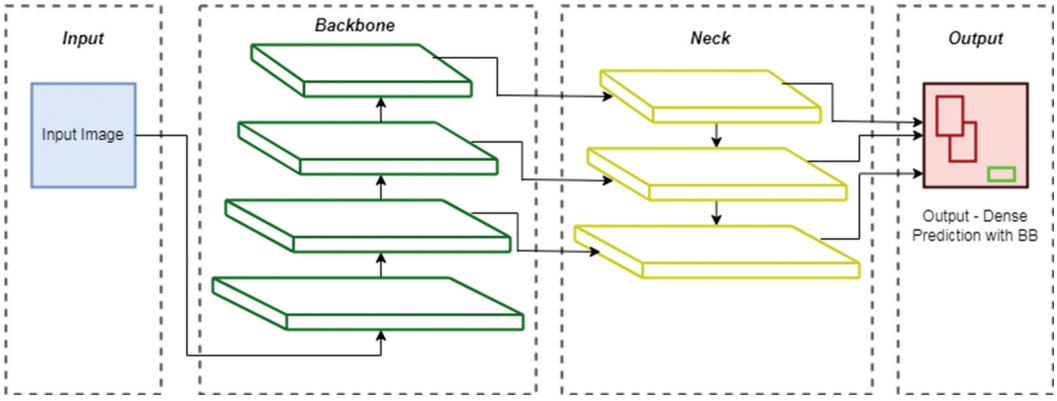

Fig. 4. YOLO: Single-stage architecture [22]

The goal of YOLO (You Only Look Once) v4 is to build a fast-functioning object detector that is also equipped for parallel processing for production systems. It had to be better in a variety of aspects, as compared to the present practices [22]. It is super-fast, high quality, and provides convincing results for object detection in terms of accuracy. Item detectors usually consist of several components: Input-This is where the picture is entered. Next, variants of Resnet-50, VGG16, ResNext50, or Darknet52, can be the backbone, which refers to the network that takes the image as input and retrieves the feature map. The neck and head are backbone sub-sets that maximize the discriminability and robustness of the function using FPN, PAN, RFB, etc., and the forecast-managing head [23-25]. For a single-stage detector such as YOLO and SSD, this may either be used for dense prediction or FRCNN and Mask RCNN, a two-stage detector also known as Sparse Prediction.

The option of architecture is the mechanism that can conjure up a suitable entity detector to be created. For the backbone, based on theoretical logic, there was an alternative between CSPResNeXt50, CSPDarknet53, and EfficientNet B3, and several CSPDarknet53 neural network tests were found to be the most optimal model. The YOLOv4 concept, the CSPDarknet53 Spatial Pyramid Pooling block, also known as SPP, was used. Since the receptive region is significantly enhanced, it distinguishes the most important context characteristics and produces practically no decrease in the speed of network traffic [26]. As a form of parameter aggregation for various detector levels from different backbone levels, PANet, also known as the Path aggregation network, is also used, and this was used instead of the FPN, also known as the YOLOv3 Feature Pyramid Network. Eventually, they chose the head of YOLOv3, as YOLOv4's head. Different classifier training features Different training features of the detector Different backbones and pre-trained training weightings of the detector Different minibatch sizes of detector training Different training features of the detector since there are several features that they had to test, particularly in the bag of freebies and specials [27]. So the approach they used was to use a methodology called ablation

analysis to test every feature. An ablation analysis is when you manually remove parts of the input to see which parts of the input are relevant to the network performance.

Normally, it looks like a table like this with the observations on the right-hand side. Speaking of results, if we look at how YOLOv4 relates to others, you would be very impressed. But to ensure that we compare each other with oranges and apples. Depicts the steps involved in the object detection process in the YOLOv4. Separate GPU architectures are used for inference cycle checking to test broadly accepted GPU architectures as competing models. The comparative GPU architectures used were the Maxwell, Pascal, and Volta architectures. You can see from these tables that YOLOv4 is comparable to the fastest and most reliable in terms of both speed and accuracy [30]. This analysis uses a state-of-the-art detector that on MS COCO datasets is faster in terms of Frames per Second (FPS) and more reliable than all available alternative detectors. On a traditional 8-16GB VRAM GPU that is readily accessible, YOLOv4 can be educated and used. The YOLOv4 is checked with a broad variety of features and the best ones are selected to improve both the classifier and the detector efficiency.

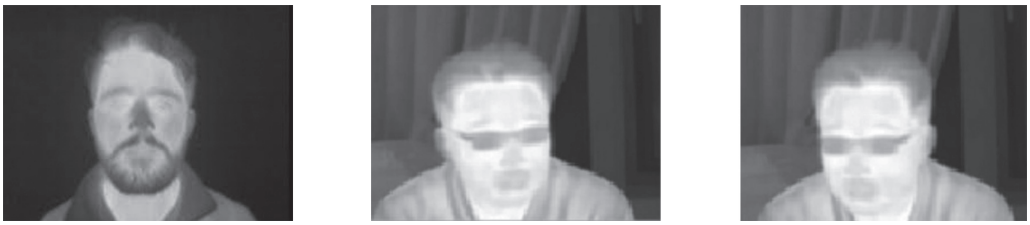

Fig. 5. Thermal images collected from RGB-D-T database

Fig. 6. The result of YOLOV4 detection in thermal images

Figure 5&6 use bounding boxes around recognized humans' faces showing the precision of thermal pictures. These detected faces are utilized to classify the face into their emotions category.

4.4. DENSENET FOR CLASSIFICATION

The classification of facial emotions in thermal images into emotions including surprise, anger, sadness, happiness, fear, disgust, and neutral. To classify human emotions from thermal face images, the DenseNet technique is used. The graphical representation of DenseNet creates a 5-layer dense block with a growth rate of k = 4 as seen in Figure 7. The 121 in DenseNet-121 stands for the total number of layers in the neural network. Combinations of different layers make up the conventional DenseNet-121 structure. It has five layers of pooling and convolution, three transition levels (6, 12, and 24), one classification layer (16), and two DenseBlocks (1x1 and 3x3 convs). The accuracy of the model for classifying the person's emotions from thermal photos is improved by DenseNet's feature reuse and parameter reduction [27]. After the composite function operation, the output of the first layer enters the second layer as an input. A non-linear activation layer, a pooling layer, a convolution layer, and a batch normalization layer make up the composite process.

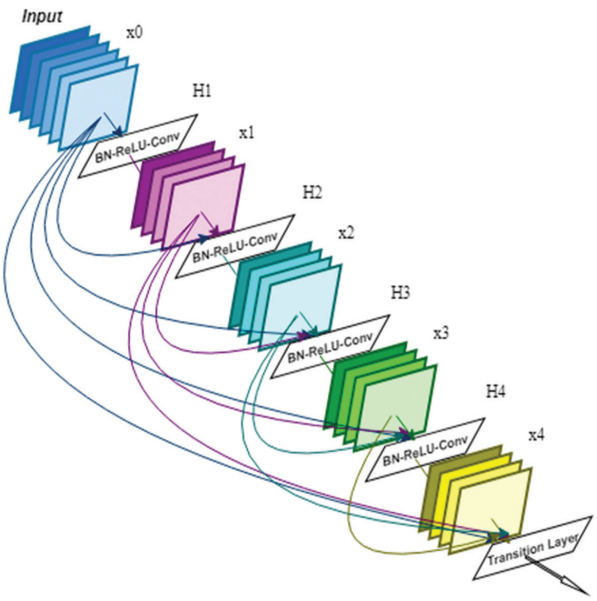

Fig. 7. Block diagram of DenseNet [27]

The Dense Block is an important part of the DenseNet for improving the information flow between layers. It is composed of BN, ReLu, and 3 × 3 Conv. The specific formula is shown as follows,

$$x_1=H_1([x_0, x_1,...,x_{l-1}]) \quad (5)$$

Where $[x_0, x_1,....,x_{l-1}]$ refers to the concatenation of the feature maps produced in layers 0, 1,…, l-1, H_l (.) is defined as a composite function of three consecutive operations on the input of l^{th} layer. The function (such as ReLU, sigmoid) to increase their nonlinearity. The convolution process can be expressed as,

$$z^l=W^l f_1 (z^{(l-1)})+b^l \quad (6)$$

Where z^l is the l^{th} -layer neuron status, f^l (.) the activation function, w^l and b^l are the weight matrix and bias from$(l\text{-}1)^{th}$ to the l^{th}, respectively. Contrary to popular assumption, DenseNets require few parameters than traditional CNNs since they do not need to know unnecessary feature maps. DenseNets layers are constrained and barely present any new feature maps. Reusing features produces highly compact versions and is the main idea behind DenseNet. Since no feature maps are duplicated, it requires fewer parameters than other CNNs. CNN encounters problems as they delve deeper. The reason for this is that the gradient in the opposite direction of the gradient from the inner layer to the outer layer gets very long that it can evaporate before hitting any farther side. By simply linking every layer to every layer, DenseNet makes this connectivity considerably simpler. By recycling features, DenseNets maximize the capacity of the network. As explained, DenseNet is a type of CNN. Typically, DenseNet architectures use dense blocks to connect all layers directly to one another, creating dense connections across layers. According to DenseNet, the more connections there are, the more accurate the system will be. In a DenseNet, each layer delivers its feature maps to the succeeding layers and receives new input from the layers above it. The idea of conjunction is utilized to describe how each layer gains collective wisdom from the levels above it. Multiple classifiers are combined into an ideal, DCNN and connected with a dense connection for effective picture categorization to maximize computation recycling between the classifiers. On the majority of them, DenseNet significantly surpasses the state-of-the-art while using the least amount of memory and processing possible to maximize its efficiency.

5. RESULTS AND DISCUSSION

This section begins by using our method to identify a face and classify their emotions on thermal pictures from the RGB-D-T database and comparing it to state-of-the-art techniques. Finally, we present the assessment findings according to experimental findings to evaluate our method in the subsections that follow.

5.1. EXPERIMENTAL DATASET

Our DenseNet model is trained using a portion of the RGB-D-T database. Thermal camera data was gathered as RGB-D-T. It has a 51-person capacity and a 384 x 288 resolution. Thermal face detection is influenced by three variables, including facial expression, head rotation, and illumination. We create the train set for our model using the thermal faces from a subset of the RGB-D-T database [28]. About 12K thermal facial photos are included. In addition, we randomly extract non-face regions in addition to the facial regions to train our model on roughly 12K photos also serving as our test set are 3K thermal photos from the database. We contrast how the various elements affect the face detection rate.

Fig. 8. Several illustrations of our thermal facial image database in various unrestricted settings.

However, the previous thermal face database is very easy to recognize when compared to the RGB face database. It is necessary to construct a database with multiple face thermal photos in unrestricted situations. RGB face database is the initial database of thermal pictures with various faces that we are aware of. The number of persons varies (from one to three), as does the head rotation (up, down, left, and right). The photographs have a resolution of 640 x 480 and were captured using 10K thermal cameras in 10 different environments. For the sake of additional investigation, we also take the relevant RGB photos. Examples from five different contexts are shown in Figure 8. Figure 8's second column features three persons with various expressions and head rotations while the first column, which was taken at night, only has one human with a neutral expression. The head rotation of the person in the third column who is leaning against the window at nighttime changes. The fourth column's spacing changes and the last column has several expressions.

5.2. EVALUATION METRICS

Each detector generates a bounding box that shows the location of people in the input photos. We can identify which predicted bounding boxes are accurate by comparing them to ground truth boxes. The intersection over union (IOU), which calculates the ratio between a pair of boxes' union and intersection, is used to quantify overlap. A perfect disjoint alignment is 0, and a perfect alignment of the two boxes is 1. A bounding box is considered to be an accurate forecast if it overlaps the ground truth by at least 0.3. Typically, an overlap of 0.5 or greater is necessary for object detection. However, it was chosen to be a little leaner here because some of these things can be rather little. Only one prediction can be mapped to a ground truth box. A prediction is deemed accurate if it falls within a ground truth box; otherwise, it is deemed incorrect.

In terms of performance measures, looked at the proposed method's Accuracy (A), Precision (P), F1-score (F), and Recall (R). These metrics indicate:

5.2.1. Accuracy

The accuracy metric is calculated to determine whether the categorization of advertisements is accurate.

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN} \quad (7)$$

5.2.2 Precision

The proportion of accurately predicted positive outcomes to all predicted positive observations is known as precision. The ability to carry out the following actions is precision:

$$Precision = \frac{TP}{TP+FP} \quad (8)$$

5.2.3. Recall

The terms for the recall are the True Positive Rate (TPR) and Sensitivity. The classifier's capacity to identify all positive samples is shown by the recall score. It is the total divided by TP, including FN. It can be described in the following terms:

$$Recall = \frac{TP}{TP+FN} \quad (9)$$

5.2.4. F-Measure

F-Measure determines the harmonic mean of recall and precision.

$$F1\ Score = 2 \times \frac{precision \times recall}{precision+recall} \quad (10)$$

5.3 QUANTITATIVE EVALUATION

In this paper, we proposed a new deep learning classifier, the DenseNet technique. Our proposed method includes four steps, pre-processing, feature extraction, detection, and classification. In Pre-processing step, we crop the input image using the DOG filter to reduce lighting variations and then normalize the image utilizing the median filter. After pre-processing, the features like the shape and location of the face are extracted from input images using by EfficientNet. The YOLOv4 method is used to detect the human face in a single frame. Then classify the facial emotion from detected face images with the help of extracted features into seven categories using the DenseNet technique. Here, we provide some thermal images which are collected from the database to detect the faces in a single thermal image and also classify the face by their emotions.

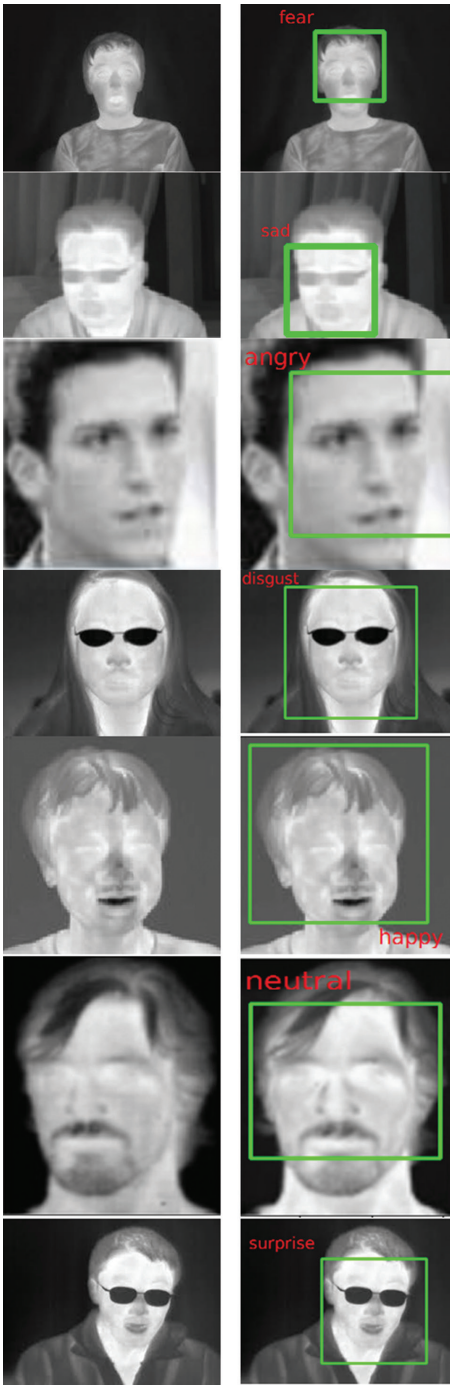

Fig. 9. The classification outcome of facial emotions using the DenseNet technique

The classified emotions on faces show in figure 9. This is classified into seven emotions.

5.4 PERFORMANCE METRICS

We compare various approaches with the RGB-D-T database studies. Table 2 shows that no matter the circumstances, our technique has the highest effectiveness and increases the recognition rate. We have no trouble identifying a single heat face opposing a dark foreground, without a doubt. Part of the reason is that the faces are closer to one another than in the photographs that featured three persons. When only one person was present in the test photographs, our approach was still able to identify the thermal face regardless of the head's rotation, occlusion, or facial expression. The proposed approach operates nicely in our database, according to the results. Table 2 lists our findings based on accuracy, precision, recall, and computing time. Compared with other techniques to classify emotions, our proposed method gives higher classification accuracy with less computation time. It can classify emotions clearly and correctly.

Table 2. Using the proposed and compared approaches, calculate precision, F-Score, accuracy, and recall (%) while classifying the computing time.

	Dataset	Precision	F-Score	Accuracy	Recall	Computing Time (ms)
SSD [14]	AAU PP T	92.04	92.5	87.46	91.24	0.15
DCNN [15]	IR Database	75.60	68.47	81.16	63.28	0.23
Faster R-CNN [16]	NVIE	89.52	90.85	93.5	93.03	0.31
YOLO V3 [17]	NVIE and PUJ	84.02	84.49	89.27	85.98	0.28
Proposed (DenseNet)	RGB-D-T database	96.15	95.34	95.97	97.42	0.16

Table 2 lists the results for SSD [14], Faster R-CNN [16], DCNN [15], YOLOV3 [17], and the planned DenseNet on the RGB-D-T database in terms of Recall, Accuracy, F-Score, and Precision. Based on the results, we can see that the proposed methodology has higher classification accuracy values than other deep learning approaches in terms of recognition rate Recall, Precision, and F-Score. The achieved F-Score for the proposed technique is 95.34%, compared to 90.85% for faster R-CNN [16], 84.49% for YOLOv3 [17], 68.47% for DCNN [15], and 92.50% for SSD [14]. The proposed method's acquired Precision is quite similar to SSD [14]'s precision. Additionally, when compared to other existing methodologies, the proposed approach's recall is superior. DCNN [15] has the lowest accuracy rate, at 81.16 percent. The better results are achieved by conducting the training phase with 26 epochs and a learning rate of 0.01. For each epoch, the experiment analysis used 31 iterations.

Figures 10-14 show how the DenseNet improves the classification outcomes. It completes the task in 0.16 milliseconds with an F-score of 95.34 percent, recall of 96.64 percent, precision of 95.92 percent, and accuracy of 95.37 percent.

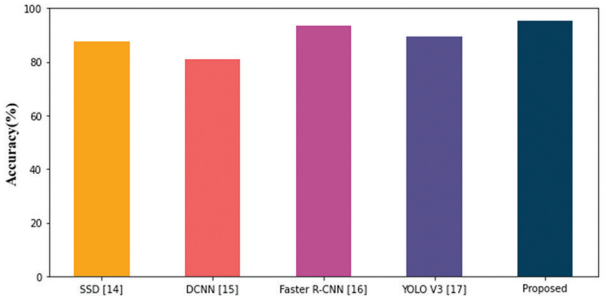

Fig. 10. Comparison of the suggested technique's classification Results with previous Methods in Terms of Accuracy

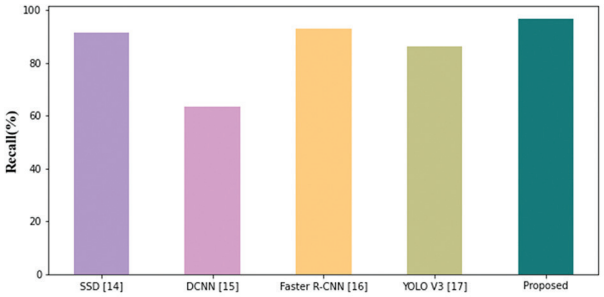

Fig. 11. Performance comparison of the classification with known methods in terms of Recall

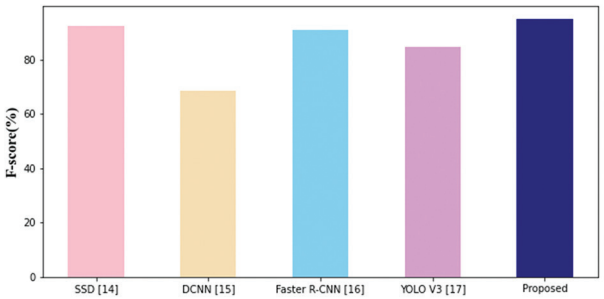

Fig. 12. Performance comparison of the classification with known methods in terms of F-score

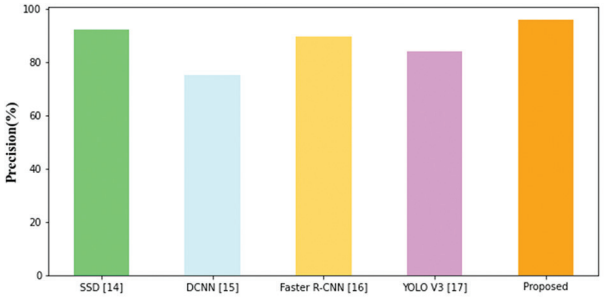

Fig. 13. Performance comparison of the classification with known methods in terms of Precision

Calculation time is another factor that is compared. Deep learning methods aim to lessen the complexity of computation. Fig. 14 displays how long the cutting-edge methods and the proposed model needed to compute using the RGB-D-T database.

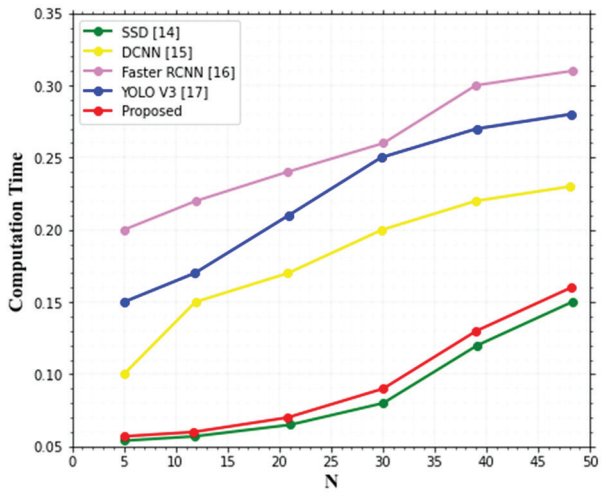

Fig. 14. Comparing the time complexity of the suggested approach to the existing techniques

From Figs. 10-14, it can be shown that the proposed strategy exceeded the other techniques and displayed the highest Accuracy, F-Score, Recall, and Precision with less consumption time.

Table 3. Human face emotions accuracy

Face Emotions	Accuracy (%)
Happy	95.6
Sad	96.9
Disgust	93.4
Neutral	99.5
Anger	97.8
Fear	98.2
Surprise	98.8

Various human facial emotions accuracy can be depicted in table 3. The various person faces emotions like happiness, sadness, surprise, anger, disgust, fear, and neutrality. Happy can yield 95.6% accuracy, sad obtain 96.9% of accuracy, disgust gain 93.4% of accuracy, neutral yield 99.5% of accuracy, anger has 97.8% of accuracy, fear obtains 98.2% accuracy and surprise gain 98.8% of accuracy. Among the seven human face emotion accuracy, the neutral emotion achieves greater accuracy. Figure 15 shows the graphical representation of the human face emotion classification.

Emotions like anger, fear, happiness, surprise, sadness, disgust, and neutrality are the ones that the model was having less accurate precision scores which are shown in Fig. 16. The most used technique for assessing classification errors is the confusion matrix. Based on the provided confusion matrix explanations, developed the confusion matrix for the DenseNet proposed model. The diagram shows that the DenseNet model can classify facial emotions with the RGB-D-T database having anger, fear, happiness, surprise, sadness, disgust, and neutral images. This shows that the proper categorization of the two statuses has been carried out. The obtained confusion matrix for the cross-validation test of classification is shown in Figure 16. Our proposed approach provides higher classification accuracy with less computation times as compared to previous techniques for classifying emotions.

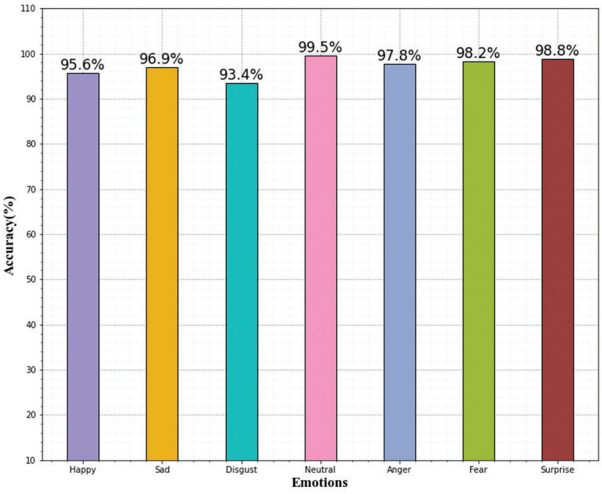

Fig.15. Evaluation of the classifier's performance in analyzing emotions in terms of Accuracy

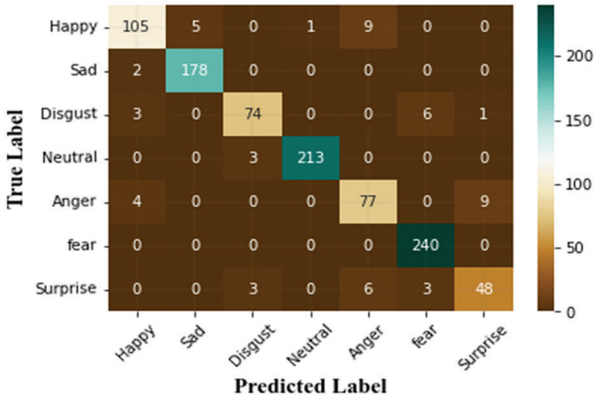

Fig. 16. Confusion matrix of facial emotions

6. CONCLUSION

In this paper, a novel DenseNet technique is proposed and introduced to thermal face emotion identification. Initially, the Difference of the Gaussian filter is used to crop the input images and then the median filter is used to normalize the input images in pre-processing step. An EfficientNet technique is used to extract the multi-scale features. The YOLOv4 technique is used for detecting the human face. Then the proposed model DenseNet is used to classify the images by extracted features. Our model performs optimally for thermal facial emotions classification, according to

the RGB-D-T results. This experiment gives 95.97% classification accuracy using the DenseNet classification technique. Our algorithm continues to perform well in general. In the future, we will gradually improve the model and further improve the detection accuracy of this algorithm based on considering the improvement of high-level features.

Conflict of interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We declare that this manuscript is original, has not been published before and is not currently being considered for publication elsewhere.

Availability of data and material: Not applicable

7. REFERENCES

[1] Y. Bi, M. Lv, Y. Wei, N. Guan, W. Yi, "Multi-feature fusion for thermal face recognition", Infrared Physics & Technology, Vol. 77, 2016, pp. 366-374.

[2] P. Saha, D. Bhattacharjee, B. K. De, M. Nasipuri, "Characterization and recognition of mixed emotional expressions in thermal face image", Infrared imaging systems: Design, analysis, modeling, and testing xxvii, International Society for Optics and Photonics, Vol. 9820, 2016, p. 98200S.

[3] Y. M. Elbarawy, R. S. El-Sayed, N. I. Ghali, "Local entropy and standard deviation for facial expressions recognition in thermal imaging", Bulletin of Electrical Engineering and Informatics, Vol. 7, No. 4, 2018, pp. 580-586.

[4] A. Sancen-Plaza et al. "Facial recognition for drunk people using thermal imaging", Mathematical Problems in Engineering, Vol. 2020, 2020.

[5] P. Saha, D. Bhattacharjee, B. K. De, M. Nasipuri, "A Thermal Blended Facial Expression Analysis and Recognition System Using Deformed Thermal Facial Areas", International Journal of Image and Graphics, 2021, p. 2250049.

[6] U. Atila, M. Uçar, K. Akyol, E. Uçar, "Plant leaf disease classification using EfficientNet deep learning model", Ecological Informatics, Vol. 61, 2021, p. 101182.

[7] Y. H. Lai et al. "Data fusion analysis for attention-deficit hyperactivity disorder emotion recognition with thermal image and Internet of Things devices", Software: Practice and Experience, Vol. 51, No. 3, 2021, pp. 595-606.

[8] A. K. Prabhakaran, J. J. Nair, S. Sarath, "Thermal facial expression recognition using modified resnet152", Advances in Computing and Network Communications, Springer, Singapore, 2021, pp. 389-396.

[9] A. I. Middya, B. Nag, S. Roy, "Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities", Knowledge-Based Systems, Vol. 244, 2022, p. 108580.

[10] D. Jiang et al. "A probability and integrated learning based classification algorithm for high-level human emotion recognition problems", Measurement, Vol. 150, 2020, p. 107049.

[11] A. Bhattacharyya, S. Saha, S. Sen, S. Mirjalili, R. Sarkar, "Deep Feature Selection Using Moth-Flame Optimization for Facial Expression Recognition from Thermal Images", Handbook of Moth-Flame Optimization Algorithm, CRC Press, 2022, pp. 281-312.

[12] M. K. Chowdary, T. N. Nguyen, D. J. Hemanth, "Deep learning-based facial emotion recognition for human–computer interaction applications", Neural Computing and Applications, 2021, pp. 1-18.

[13] S. K. Km, R. Rajendran, Q. Wan, K. Panetta, S. S. Agaian, "TERNet: A deep learning approach for thermal face emotion recognition", Mobile Multimedia/Image Processing, Security, and Applications, International Society for Optics and Photonics, Vol. 10993, 2019, p. 1099309.

[14] K. R. Akshatha et al. "Human Detection in Aerial Thermal Images Using Faster R-CNN and SSD Algorithms", Electronics, Vol. 11, No. 7, 2022, p. 1151.

[15] A. Bhattacharyya, S. Chatterjee, S. Sen, A. Sinitca, D. Kaplun, R. Sarkar, "A deep learning model for classifying human facial expressions from infrared thermal images", Scientific Reports, Vol. 11, No. 1, 2021, pp. 1-17.

[16] S. Nayak, B. Nagesh, A. Routray, M. A. Sarma, "Human–Computer Interaction framework for emotion recognition through time-series thermal video sequences", Computers & Electrical Engineering, Vol. 93, 2021, p. 107280.

[17] N. K. Benamara, E. Zigh, T. B. Stambouli, M. Keche, "Towards a Robust Thermal-Visible Heterogeneous Face Recognition Approach Based on a Cycle Generative Adversarial Network", International Journal of Interactive Multimedia and Artificial Intelligence, Vol. 7, No. 4, 2022.

[18] M. F. Siddiqui, A. Y. Javaid, "A multimodal facial emotion recognition framework through the fusion of speech with visible and infrared images", Multimodal Technologies and Interaction, Vol. 4, No. 3, 2020, p. 46.

[19] S. Pachade, P. Porwal, M. Kokare, L. Giancardo, F. Mériaudeau, "NENet: Nested EfficientNet and adversarial learning for joint optic disc and cup segmentation", Medical Image Analysis, Vol. 74, 2021, p. 102253.

[20] T. Liu, B. Pang, L. Zhang, W. Yang, X. Sun, "Sea Surface Object Detection Algorithm Based on YOLO v4 Fused with Reverse Depthwise Separable Convolution (RDSC) for USV", Journal of Marine Science and Engineering, Vol. 9, No. 7, 2021, p. 753.

[21] I. Sim, J. H. Lim, Y. W. Jang, J. You, S. Oh, Y. K. Kim, "Developing a Compressed Object Detection Model based on YOLOv4 for Deployment on Embedded GPU Platform of Autonomous System", arXiv:2108.00392, 2021.

[22] J. H. Sejr, P. Schneider-Kamp, N. Ayoub, "Surrogate Object Detection Explainer (SODEx) with YOLOv4 and LIME", Machine Learning and Knowledge Extraction, Vol. 3, No. 3, 2021, pp. 662-671.

[23] X. Sun, T. Liu, X. Yu, B. Pang, "Unmanned Surface Vessel Visual Object Detection Under All-Weather Conditions with Optimized Feature Fusion Network in YOLOv4", Journal of Intelligent & Robotic Systems, Vol. 103, No. 3, 2021, pp. 1-16.

[24] D. Wu, S. Lv, M. Jiang, H. Song, "Using channel pruning-based YOLO v4 deep learning algorithm for the real-time and accurate detection of apple flowers in natural environments", Computers and Electronics in Agriculture, Vol. 178, 2020, p. 105742.

[25] N. Kumari, V. Ruf, S. Mukhametov, A. Schmidt, J. Kuhn, S. Küchemann, "Mobile Eye-Tracking Data Analysis Using Object Detection via YOLO v4", Sensors, Vol. 21, No. 22, 2021, p. 7668.

[26] M. Zhang, S. Xu, W. Song, Q. He, Q. Wei, "Lightweight Underwater Object Detection Based on YOLO v4 and Multi-Scale Attentional Feature Fusion", Remote Sensing, Vol. 13, No. 22, 2021, p. 4706.

[27] X. Li, X. Shen, Y. Zhou, X. Wang, T. Q. Li, "Classification of breast cancer histopathological images using interleaved DenseNet with SENet (IDSNet)", PloS One, Vol. 15, No. 5, 2020, p. e0232127.

[28] D. Mahouachi, M. A. Akhloufi, "Adaptive deep convolutional neural network for thermal face recognition", Thermosense: Thermal Infrared Applications XLIII, Vol. 11743, SPIE, 2021, pp. 15-22.