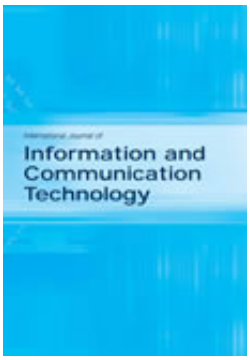
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

Resilient recognition system for degraded thermal images using convolutional neural networks

Naser Zaeri, Rusul R. Qasim

Article History:

Received:	22 May 2023
Last revised:	16 May 2024
Accepted:	28 June 2024
Published online:	02 August 2024

Resilient recognition system for degraded thermal images using convolutional neural networks

Naser Zaeri*

Faculty of Computer Studies,
Arab Open University,
P.O. Box 830 Ardiya 92400, Kuwait
Email: n.zaeri@aou.edu.kw
*Corresponding author

Rusul R. Qasim

Kuwait Technical College,
P.O. Box 232, Abu-Halifa, 54753, Kuwait
Email: r.qasim@ktech.edu.kw

Abstract: For biometric identity applications, thermal infrared face recognition technologies have become a powerful alternative to visual systems. However, thermal images can undergo degradation in various ways, including noise, blurring, reduced spatial resolution, and temperature drift, in addition to being affected by changes in pose and facial expression. In this paper, we propose using convolutional neural networks (CNNs) to recognise degraded thermal face images. The system deals efficiently with poor-quality images resulting from various causes. We describe how a CNN structure processes images and use ResNet-50 architecture to demonstrate our results, being an essential deep learning model that has proven its efficiency in various computer vision and machine learning applications. We conduct experiments under different conditions and scenarios that tackle quality, reduced spatial resolution, pose, and expression variations challenges. To evaluate the performance of the proposed method, we conduct thorough experiments and detailed analysis on a database of 7,500 images. The results demonstrate that the proposed system provides greater discriminability and robustness against such variations, as well as higher identification rates under various situations reflecting real-world scenarios, compared to other recently published work.

Keywords: thermal image; face recognition; low resolution; convolutional neural networks; CNNs; pose variations.

Reference to this paper should be made as follows: Zaeri, N. and Qasim, R.R. (2024) 'Resilient recognition system for degraded thermal images using convolutional neural networks', *Int. J. Information and Communication Technology*, Vol. 25, No. 5, pp.50–71.

Biographical notes: Naser Zaeri is an Associate Professor with the Department of Information Technology and Computing (ITC) at the Arab Open University (AOU), Kuwait. He has obtained his PhD in Electrical Engineering from the University of Surrey, UK in 2008. He has obtained his MSc and BSc in Electrical Engineering from the Kuwait University (Honour List). He served as the Director of Research and Development at the AOU during (2012–2014), and was the Head of ITC Department at the AOU during (2010–2012). He

served as a consultant for many authorities and ministries and implemented many projects in various engineering and technology fields. He has more than 50 different publications in international journals and conferences. His areas of interest are digital image and signal processing, biometrics, pattern classification, communications systems, and wireless sensor networks.

Rusul R. Qasim is an instructor with the Kuwait Technical College. She received her BSc in Information Technology and Computing from the Arab Open University, Kuwait. She received her BSc in Information Technology and Computing from The Open University, UK in 2017, as well. She has obtained her MSc in Information Systems from the University of Jordan, Jordan in 2019. Her research interests include machine learning, theory of algorithms, blockchain, information systems and software engineering.

1 Introduction

Over the last 20 years, face recognition technology has significantly advanced in various domains using traditional visible-light images. However, visible-spectrum images are very variable due to the fact that they are created by surface reflection, which is highly reliant on luminance and the spatial distribution of light sources (Lezama et al., 2017). Face identification using visible images could not have the essential resilience in real-world scenarios where lighting conditions, weather, or time might alter illumination intensities, resulting in poorer identification rates. In addition, the skin tones of various ethnic groups affect how differently light is reflected from human faces. Moreover, pose variations result in substantial visual appearance changes and diminish face recognition performance. Also, automated systems suffer many difficulties, such as distinguishing faces in videos, using illegal face disguises, and facial expressions (Iranmanesh et al., 2018).

For biometrics identity applications and military enforcement, the thermal infrared (IR) facial recognition system has become a powerful alternative modality in response to the difficulties of visible facial recognition and the diminished recognition ability of algorithms caused by the factors mentioned above. IR technology captures anatomical data that contains underlying characteristics believed to be unique to each individual (Lin et al., 2019). Thermal emissivity from the facial surface is measured using IR cameras, and their images are relatively stable in light variations. Unlike visible light, which is more susceptible to dispersion and absorption by smoke and dust, IR energy may be detected in any lighting situation. Since thermal IR sensors only detect the heat pattern generated by the subject, thermal face images are independent of ambient lighting conditions. Depending on their temperature and characteristics, many substances emit various types of IR radiation. The human face and body temperature ranges are comparable and reasonably homogeneous. This ensures a balanced thermal signature. Thermal patterns on the face are formed mainly through the pattern of superficial blood vessels underneath the skin. Each individual's face's vascular, and tissue structure is unique; hence, so are the IR images (Silverthorn, 2015).

However, face recognition based on IR thermography is still a largely unexplored field since many facets of thermal IR system performance have yet to be well investigated. Besides the effects of pose and expression variations, thermal images can

experience degradation in different forms, such as noise, blurring, loss of spatial resolution, and temperature drift. Excessive noise can cause image distortion while blurring decreases image sharpness. Loss of spatial resolution can impact the level of detail in an image, while temperature drift can lead to mistakes in temperature measurements over time or due to environmental conditions. Additionally, alignment of objects and the configuration of areas in the image can impact the system performance. Other important factors include pixel size. In thermal imaging cameras, each pixel represents a particular area of the measured temperature. In cases where the captured item is small or situated at a considerable distance, the thermal camera could lack adequate pixels to produce a comprehensive image, resulting in a blurred visual representation. Further, the lens quality used in a thermal imaging system can also affect the sharpness of the image. Lower-grade lenses may generate distortions or anomalies that can cause the image to appear fuzzy. Also, like any imaging system, thermal cameras must be appropriately focused to capture clear images. An incorrect focus of the camera can result in an ambiguous image.

All of the above issues emphasise the necessity of proposing an efficient system that can deal with such challenges. In this paper, we investigate the use of CNNs in recognising thermal face images experiencing degradation in different forms. We believe that the CNN model will be able to provide a powerful recognition capability that can cope well with such images. Recently, deep learning and CNNs have made tremendous strides in solving various tasks of computer vision problems (Kumar and Singh, 2020). CNNs and deep networks extract low, middle and high-level features in an end-to-end multi-layer manner, where the number of stacked layers can enrich the 'levels' of features. They are designed to be invariant to object position and distortion in the scene, where they can achieve the complex function approximation through a nonlinear network structure. Hence, they are very good in extracting patterns in the input image, such as lines, gradients, circles, or even eyes and faces. It is this property that makes CNNs so powerful for computer vision. Unlike earlier computer vision algorithms, CNNs can operate directly on a raw image and do not need any pre-processing. In this regard, it is worthy to note that there are various factors which helped in the improvement of deep learning networks performance. The development of parallel computing frameworks and the exploitation of new approaches in the design of models' architecture and training methodologies beside the evolution of new techniques to overcome the problem of over-fitting using data augmentation and batch normalisation are a few examples of these factors. These characteristics and features made deep learning and CNNs very efficient candidates for the task in hand (Gonzalez, 2018).

In this regard, we provide a mathematical analysis to describe how a CNN structure processes images and investigate the use of one critical CNN model, the ResNet-50 architecture (He et al., 2016). The rest of the paper is described as follows. Section 2 presents recent related works. In Section 3, we discuss the methodology. Section 4 presents the analytical and experimental results. Finally, the paper is brought to conclusion in Section 5.

2 Related works

This section provides a brief overview of recent and up-to-date works addressing face recognition using thermal imaging. Maeng et al. (2012) used scale-invariant feature

transform and multi-scale local binary pattern for thermal image feature extraction. Their results indicated that the former has better accuracy compared to the latter. In Bi et al. (2016), multiple traits were merged for thermal face characterisation, including Gabor jet descriptor, local binary pattern, and down-sampling feature. According to their tests, the proposed solution is noise and occlusion resistant and outperforms systems that use a single feature. The issues caused by the temporal changes of infrared facial images were examined by Vigneau et al. (2017). The temporal variations are primarily caused by various environmental factors, physiological changes, and variations in the responsiveness of infrared detectors. Kumar and Singh (2020) proposed a CNN architecture framework for occluded images by analysing the performance of pre-trained models using transfer learning. In order to enhance the system performance, they used different decision-level fusion strategies. Their work offered better results compared to a single CNN architecture. Lin and Chen (2019) proposed a system using model fusion based on a CNN, where a grid of thermal points based on physiological data is extracted to compute the support vectors to discover the hyperplane necessary for classification.

Using the particle swarm optimisation methodology, Hermosilla et al. (2018) established a fusion method that integrates thermal and visible descriptors. Weights are applied to descriptors to prioritise particular areas of a fused image. The system uses descriptors generated by different combinations of local matching descriptors: local binary pattern, local derivative pattern, and oriented gradients histograms. Lin et al. (2020) obtained thermal image features using random forest, deep learning, and ensemble learning to build a face model. The proposed feature extraction method divides the facial image into blocks before generating the feature matrix. Pini et al. (2021) used VGG16, ResNet-18, and InceptionV3 for feature extraction, where they compared probe and gallery depth maps using cosine similarity. Authors in Hermosilla et al. (2021) proposed using generative adversarial networks to create high-quality synthetic thermal images and obtain training data to build the recognition models. They aimed at generating synthetic thermal images by training neural models using a distribution of input data in order to generate new data that resemble the original ones. Kakarwal et al. (2020) applied backpropagation and Levenberg-Marquardt algorithms on visible and thermal fused imagery. The backpropagation algorithm achieved an accuracy of 92.86%, whereas the Levenberg-Marquardt accuracy was 83.92%.

3 Methodology

As discussed in Section 1, we approach the solution to the problem by using a CNN architecture. To retain the spatial information of an image, CNNs are typically utilised in image processing as a deep learning method that uses supervised learning (Gonzalez, 2018). It comprises layers of convolution, activation, and pooling. In a CNN, tens of these steps may be linked together in a chain. Architectures for CNNs vary not just in the number of stages but in the definition and usage of components inside each. Input maps, feature maps, and pooled maps make up the three volumes that constitute a CNN stage. All maps are two-dimensional arrays whose sizes vary depending on volume; however, within a given volume, all maps are similar. Convolutional and pooling layers are added between the original CNN's input and output layers for improved data processing performance. Convolution work extracts high-level features from input data. Pooling work, like the convolutional layer, is responsible for lowering the spatial size of the

convolved feature (He et al., 2016; Wu, 2017). The computer power required to process the data is reduced through dimensionality reduction. For each input image, the CNN's output is fed into a deep, fully connected network (FCN) whose goal is to convert a set of two-dimensional features into a class label. The capacity to learn the operational parameters of each network layer using sample training data is crucial.

Let us suppose $\boldsymbol{t}$ is the corresponding target (ground-truth) value for the input $\boldsymbol{x}^1$, then a cost or loss function can be used to measure the discrepancy between the CNN prediction $\boldsymbol{x}^L$ and the target $\boldsymbol{t}$. For example, a simple loss function could be

$$z = \frac{1}{2}\|\boldsymbol{t} - \boldsymbol{x}^L\|^2 \tag{1}$$

Eventually, the second layer receives $\boldsymbol{x}^2$, and the process continues to the following layers. Finally, $\boldsymbol{x}^L \in \mathbb{R}^C$ is obtained, which estimates the posterior probabilities of $\boldsymbol{x}^1$ belonging to the C categories (classes). The CNN prediction is expressed as

$$\arg\max_i x_i^L \tag{2}$$

The model parameters are learned using stochastic gradient descent (SGD) technique. The loss z is a supervision signal, guiding how the parameters of the model should be modified (updated). The SGD modifies the parameters by

$$\boldsymbol{\omega}^i \leftarrow \boldsymbol{\omega}^i - \eta \frac{\partial z}{\partial \boldsymbol{\omega}^i} \tag{3}$$

In every update the parameters are changed by a small amount of the negative gradient, controlled by a learning rate ($\eta > 0$), usually set to 0.001. The gradient is computed using error back propagation. In a convolution layer, multiple convolution kernels are usually used. Assuming D kernels are used and each kernel is of spatial span $H \times W$, we denote all the kernels as $\boldsymbol{f}$. Thus, the convolution procedure can be expressed as

$$y_{i^{l+1}, j^{l+1}, d} = \sum_{i=0}^{H} \sum_{j=0}^{W} \sum_{d^l=0}^{D^l} f_{i,j,d^l,d} \times x^l_{i^{l+1}+i, j^{l+1}+j, d^l} \tag{4}$$

Equation (4) is repeated for all $0 \le d \le D = D^{l+1}$, and for any spatial location (i^{l+1}, j^{l+1}) satisfying $0 \le i^{l+1} < H^l - H + 1 = H^{l+1}$, $0 \le j^{l+1} < W^l - W + 1 = W^{l+1}$. Finally, rectified linear unit (ReLU) layer is used to increase the nonlinearity of the CNN. The ReLU function is a nonlinear function that maintains the input's original size, so $\boldsymbol{x}^l$ and $\boldsymbol{y}$ have similar sizes. One way to think of it is as a separate truncation for each element in the input

$$y_{i,j,d} = \max\left\{0, x^l_{i,j,d}\right\} \tag{5}$$

with $0 \le i < H^l = H^{l+1}$, $0 \le j < W^l = W^{l+1}$, and $0 \le d < D^l = D^{l+1}$. A more detailed discussion about the above analysis can be found in Wu (2017) and Zaeri and Qasim (2023). Pooling, or subsampling, is essentially a lower resolution feature map. When pooling data, it is a usual practice to use the average value of each neighbourhood to replace the original values in the feature maps. Compared to the feature maps, the pooled maps that result from using a 2×2 neighbourhood are half as large in each spatial dimension. As a result of pooling, substantial data reduction occurs, which aids in processing speed. The design of a CNN is affected, in a manner analogous to that of activation functions, by the

type of pooling that is used. In addition to the method of *neighbourhood averaging*, another method of pooling is known as *max pooling*. This method replaces the values in a neighbourhood with the highest value among its members. Max pooling has been shown to be specifically successful at classifying large image datasets, with the added benefit of speed and simplicity.

Feature maps amount present at each stage of a CNN (as well as whether or not pooling occurs at that level) define the fundamental architecture of that stage. Indicated as well are the kernel size and the pooling size, in addition to the convolution stride, which can be regarded as the number of incremental shifts in the kernel position that occur between each convolution operation. For example, for a stride of two, the convolution operation will be performed at every other spatial location in the input maps. The output maps are then passed into an FCN, aiming to categorise the input into one of a predetermined number of classes. This occurs during the last stage of a CNN. An FCN comprises layers of units known as artificial neurons, with each neuron's output in one layer coupled to the input of every neuron in the following layer. The number of neurons in the output layer equals the number of pattern classes in a specific application.

Actually, the architecture of a CNN can vary based on the specific requirements of the recognition task and the characteristics of the input image. These requirements and characteristics can be explained as follows:

- *Size of the input image:* the dimensions of the input image (e.g., width, height, colour channels) influence the choice of network architecture, particularly the size and number of layers.
- *Complexity of the recognition task:* tasks like image classification, object detection, segmentation, etc., may require different architectures. For instance, more complex tasks might need deeper networks with additional layers.
- *Texture and patterns:* the presence of specific textures, patterns, or structures in the image, which may require the network to have specialised layers or filters to detect them effectively.
- *Amount of available data:* the size and quality of the available dataset can determine the depth and complexity of the network. More data might allow for more extensive networks with more parameters.
- *Performance requirements:* the desired accuracy, speed, and efficiency level can influence the architecture choice. For real-time applications, a lighter architecture might be preferred.
- *Computational resources:* the hardware available for training and inference, such as CPU capabilities, can impact the choice of architecture. Larger networks require more computational resources.
- *Memory constraints:* in scenarios where memory usage is a concern, such as deploying models on mobile devices, architectures with fewer parameters and lower memory footprint are favoured.
- *Robustness:* the ability of the network to perform well under various conditions, such as changes in lighting, scale, rotation, and occlusion.

- *Scalability:* the ability of the network to handle larger datasets or more complex tasks without a significant decrease in performance.
- *Generalisation:* the capability of the network to perform well on unseen data, indicating its ability to learn meaningful features and patterns rather than memorising specific examples.

Researchers have recently proposed different sets of architectures (Hermosilla et al., 2018). In this work, we adopt the ResNet-50 (He et al., 2016), which has proven its efficiency in dealing with vast amounts of data with varying complexities. It offers several benefits and advantages. From a depth point-of-view, it is deeper compared to other CNN architectures such as VGG or AlexNet. Its depth allows it to learn more complex features, leading to better performance in tasks such as image classification and object detection, especially in harsh applications like ours. Moreover, ResNet-50 introduces residual connections, which help address the vanishing gradient problem. Allowing the network to learn residual functions makes it easier for the network to propagate gradients during training, enabling the training of deeper networks without suffering from degradation in performance. This specific advantage further enables more accessible training of deeper networks. With residual learning, the network can effectively learn from both the identity mapping and the residual mapping, which facilitates smoother training and faster convergence. Further, the proposed method maintains computational efficiency despite its depth due to residual connections, which reduce the number of parameters and computational costs compared to an FCN of similar depth. These benefits contribute to its ability to learn highly discriminative features that compensate for the low-quality nature of the thermal image we are dealing with. This eventually will lead to excellent performance on challenging datasets.

In more detail, ResNet-50 is a 50-layer CNN that forms a network by stacking residual blocks, where each individual weight layer is implemented as a 3×3 convolution. There are no FCN layers until the final layer. Residual networks use a skip connection or a 'shortcut' between every two layers along with using direct connections among all the layers. This allows it to take activation from one layer and feed it to another layer, hence sustaining the learning parameters of the network in deeper layers. An important effect of using the residual blocks is that they solve the 'vanishing gradient' problem, where the gradient signal diminishes in layers that are farther away from the end of the network. Eventually, the network learns the difference between the input and output, and the overall accuracy is increased. In other words, all the layers in the network will always produce the optimal feature maps, that is; the best case feature map after the convolution, pooling and activation operations. This optimal feature map contains all the pertinent features that can classify the image to its ground-truth class.

Since ResNet-50 uses 1 × 1 convolutions, the number of parameters and matrix multiplications are reduced. This enables much faster training of each layer. This architecture not only decreases the complexity of the model, but it is also efficient since it can deal with repeated patterns effectively. Moreover, this architecture can detect and capture the complex and global patterns in images, while maintaining a low error rate. To summarise, the ResNet-50 consists of 50 layers:

- Convolutional layers:
  a The majority of the layers in ResNet-50 are convolutional layers. These layers apply convolutional filters to extract features from the input image.

   b The activation function commonly used in these layers is the ReLU.
   c ReLU helps introduce nonlinearity into the network and allows it to learn complex representations.

- Pooling layers:
   a ResNet-50 includes average pooling/max pooling layers.
   b These layers do not have an explicit activation function; they simply aggregate information from the previous layer.

- Fully connected layer:
   a The final layer of ResNet-50.
   b Uses the softmax function as the activation function which converts raw scores into class probabilities.

4 Experimental analysis and results

Due to the immaturity of the work using thermal images, researchers have no well-constructed framework or consensus about the best thermal database structure, especially when it comes to the definition of 'degraded images'. A thermal image database should depict individuals with varying facial expressions and poses. Additionally, the images must be captured throughout time under practical life scenarios. In order to compare our work with a ground-truth reference that targets similar circumstances and goals, we have utilised the original dataset used in Zaeri (2020) to implement the proposed technique. The dataset was built using a thermal imaging system that employs a micro-bolometer with an image resolution of 320 × 240 focal plane array.

Subjects were instructed to exhibit three distinct expressions, where Expression 1 indicates a 'neutral' expression, Expression 2 represents an 'angry' expression, and Expression 3 represents a 'happy' expression. Five images were captured at five different angles for each facial expression: 0°, 45°, 90°, 135°, and 180°. The image at 0° portrays a person gazing over his/her right shoulder. In addition, an image at 90º reflects the frontal pose, while an image at 180° depicts a person gazing over his/her left shoulder. Twenty distinct persons have participated in image capture sessions, and each is characterised by 75 different images (5 images × 5 poses × 3 expressions). This generates a dataset of 1,500 images. All the images were cropped to size 180 × 160. We refer to this dataset as the original image dataset. Figure 1 illustrates thermal images of a few participants in five positions. It should be highlighted that the dataset includes images of individuals wearing eyeglasses. This creates an additional barrier for the dataset, as eyeglasses block a sizeable portion of the thermal radiation emitted from the eyes' area. In addition, acquisitions occurred at varying stages across a few months.

Furthermore, these original images are down-sampled to smaller sizes to obtain lower-resolution versions of the dataset. More precisely, the images are down-sampled to 90 × 80, 45 × 40, 22 × 20, and 18 × 16. Eventually, we end up with 7,500 images (where for each of the five different resolutions we have 1,500 images). Testing the proposed method on such a vast and challenging dataset should demonstrate the rigidity performance of the system and prove its efficiency. In our experiments, a 90 × 80 image is denoted as an image of 0.5-resolution (since the original image has been reduced by

half (50%) of its rows and columns). Similarly, the 45 × 40, 22 × 20, and 18 × 16 are denoted as quarter (0.25), one-eighth (0.125), and one-tenth (0.1), respectively. Figure 2 shows examples of these images for one subject at the five poses.

Figure 1 Examples from the thermal image dataset showing different expressions and poses

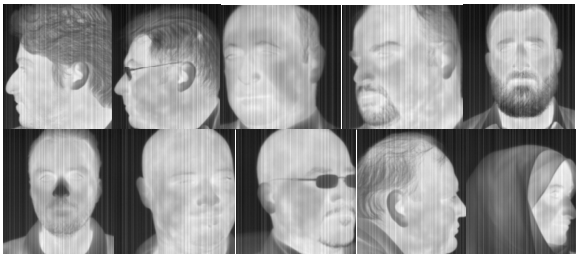
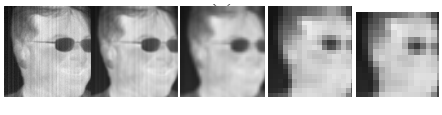


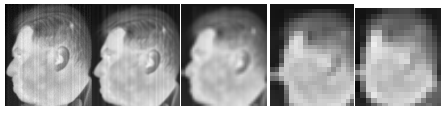
Figure 2 Examples of original (left-most image), 0.5, 0.25, 0.125, and 0.1 (right-most image) resolution for one participant: (a) frontal images, (b) pose = 45°, (c) pose = 135°, (d) pose = 0°, (e) pose = 180°

(a)

(b)



(c)



(d)

(e)

We have implemented different experiments to investigate the system's performance. We started with the original dataset to find how the system responds to changing the training set size. We implemented various training scenarios (hence, experiments), as in Table 1. The images used in the training sets are randomly chosen for each training set. For example, for the 10% training set size, 150 images are randomly selected from the 1,500

images, and the rest 1,350 images are used for testing. Every training scenario is executed for five epochs. In artificial neural networks, an epoch refers to a cycle over the whole training dataset. Typically, training a neural network requires many epochs.

Table 1 Training scenarios for a single dataset

Percentage (%)	*5*	*10*	*15*	*20*	*30*	*40*	*50*	*60*	*70*	*80*
Corresponding number of training images	75	150	225	300	450	600	750	900	1,050	1,200

Figure 3 shows the system accuracy (correct recognition rate) versus the training set size. The figure shows that the system has a correct recognition rate of almost 60% when trained using only 5% of the dataset. However, this rate increases remarkably by only utilising 10% of the dataset to achieve 84.5% and continues to increase to 98.1% when 20% of the dataset is used. Eventually, the system's accuracy reaches 100% when the system uses 60% of the dataset for training (after five epochs). The above experiments have been repeated for the other datasets of lower resolution. Figure 4 shows the system accuracy versus the training set size corresponding to the 0.5-resolution dataset. The figure reveals that the performance measures remain almost identical to the original dataset case. Figure 5 shows the system accuracy for the 0.25-resolution dataset. This figure shows that the system presents a correct recognition rate of 50% when trained on 5% of the dataset. This rate, however, increases to 81% when 10% of the dataset is used and keeps increasing to 97% when 20% of the dataset is used. Ultimately, the system's accuracy reaches 100% when using 60% of the dataset for training (again after five epochs). Almost the same observations can be noticed when examining Figures 6 and 7, corresponding to 0.125-resolution and 0.1-resolution datasets, respectively. The accuracy of the system for these two datasets saturates at 99.6%.

To get a deeper insight into system performance, we present studies on the accuracy rate and the corresponding loss versus the number of epochs. Due to space limitations, we only present the results for the training set size of 20%. Figure 8 shows system accuracy versus the number of epochs for the original dataset case, whereas Figure 9 demonstrates the corresponding loss. As can be noticed from these two figures, the face recognition accuracy reaches 98% by epoch 4. The system maintains the same performance for the cases of 0.5-resolution and 0.25-resolution, as seen in Figures 10–13. The accuracy slightly lowers down for 0.125-resolution and 0.1-resolution, reaching 96% and 94% (by epoch 4), respectively, as shown in Figures 14–17.

Further, we assess the performance of the system using the confusion matrix. The confusion matrix is one of the critical indicators suggested in recent years to better understand the system's performance (Ruuska et al., 2018). A confusion matrix is found by calculating the number of correctly identified class data (true positives), the number of correctly identified data that do not belong to the class (true negatives), and data that were incorrectly assigned to the class (false positives) or that were not recognised as class data (false negatives). Hence, a confusion matrix for k-class classification is a $k \times k$ contingency table whose cells $[i, j]$ ($i = 1, \ldots, k$, $j = 1, \ldots, k$) present frequencies of observations with real class C_i and inferred class C_j. Confusion matrix analysis has various benefits, including resistance to data distribution and relationship type. Additionally, it gives a thorough evaluation of validity and more details on the many kinds and causes of errors. Figure 18 shows the resulting confusion matrix for the original dataset corresponding to the case of 20% training set size. As Figure 18 reveals,

only a few images from certain classes have been misclassified as other classes. The number of the misclassified images marginally changes in other datasets, as can be observed in Figures 19–22, which show the confusion matrices for the 0.5-resolution, 0.25-resolution, 0.125-resolution, and 0.1-resolution datasets, respectively.

Furthermore, the system performance is evaluated using other metrics including *recall* (*sensitivity*), *specificity*, *precision*, and *F-score*. *Recall* is defined as the number of properly recognised positive examples divided by the number of positive examples in the dataset. High *recall* indicates that the model is good at identifying positive instances from the total actual positive instances in the dataset. *Specificity* illustrates how successfully a classifier recognises negative labels, while *Precision* gives the number of correctly classified positive cases divided by the number of all instances labelled by the system as positive. High *precision* indicates that when the model predicts a positive result, it is likely to be correct. The *F-score* provides a combination of the recall and precision metrics. Precision and recall are crucial aspects of a model's performance, but they can sometimes be at odds. For instance, increasing precision may lower recall and vice versa. The *F-score* is a measure that balances both precision and recall. It is the harmonic mean of precision and recall, providing a score that considers false positives and false negatives. In many real-world scenarios, the classes in the dataset are imbalanced, meaning one class has significantly more instances than the other(s). In such cases, accuracy alone can be misleading, as a model could achieve high accuracy by predicting the majority class most of the time.

In some applications, false positives and false negatives have different costs or consequences. For example, in medical diagnosis, a false negative (failing to diagnose a disease) can have more severe consequences than a false positive (incorrectly diagnosing a healthy person). By incorporating both precision and recall, the *F-score* provides a holistic measure of performance that considers the costs associated with different types of errors as it helps to give a complete picture of how well a model is performing, making it easier to compare different models or tune hyperparameters. Equations (6)–(9) represent the aforementioned metrics.

$$\text{Recall (Sensitivity)} = \frac{TP}{TP + FN} \tag{6}$$

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{7}$$

$$\text{Precision} = \frac{TP}{TP + FP} \tag{8}$$

$$\text{F-score} = \frac{2TP}{2TP + FP + FN} \tag{9}$$

where TP, TN, FP, and FN values are the number of true positives, true negatives, false positives, and false negatives, respectively. Actually, we have implemented an exciting and rigorous experiment to test the metrics represented in equations (6)–(9). We combined all datasets of different resolutions to form one colossal dataset of all 7,500 images and found *recall, specificity, precision*, and *F-score* for different training datasets. Moreover, the training set is taken randomly from all resolutions for each class. As such, the training set will have various numbers from different resolutions. Eventually, the rest

of the images for the corresponding class are used for testing. Figures 23–26 show the corresponding results for different training sets. As can be deduced from Figure 23, the system 'recall' is 0.88 when 5% of the set is used for training. This value increases even more and eventually reaches 0.99 when only 30% of the set is used for training. Regarding the 'specificity', we can see from Figure 24 that the results are very solid and range between 0.99 to 1. The same conclusion can also be concluded from Figures 25 and 26 when analysing the system's 'precision' and 'F-score'. Finally, Table 2 illustrates the performance of baseline systems that employ the infrared spectrum for applications involving the recognition of faces. The featured publications examine diverse methodologies and approaches employed by numerous research groups. Table 2 clearly demonstrates that the proposed method attains exceptional outcomes. The findings indicate that this method has balanced performance and is robust when handling thermal images of low quality resulting from various causes.

Figure 3 System accuracy versus training set size corresponding to the original dataset (see online version for colours)

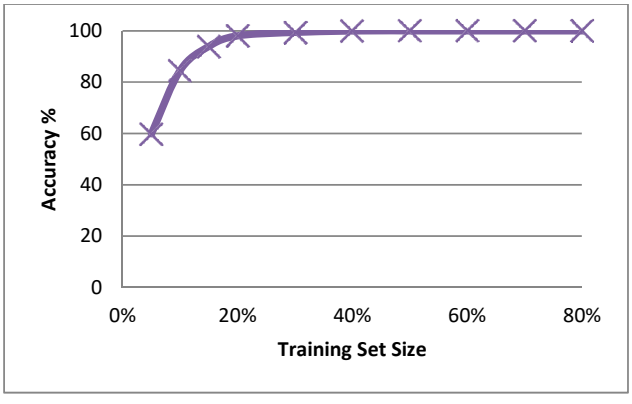


Figure 4 System accuracy versus training set size corresponding to the 0.5-resolution dataset (see online version for colours)

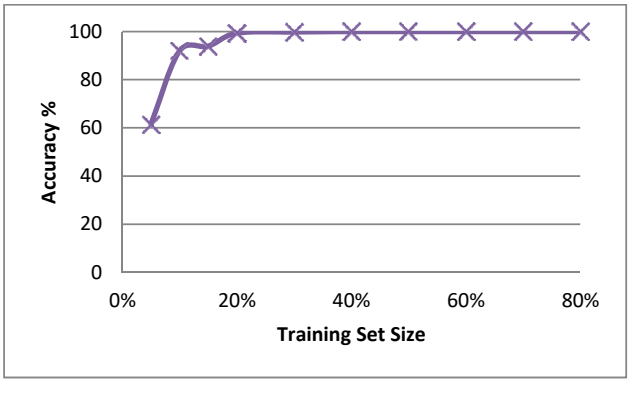


Figure 5 System accuracy versus training set size corresponding to the 0.25-resolution dataset (see online version for colours)

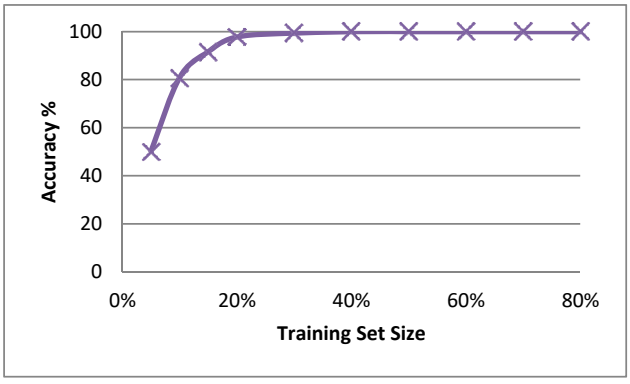


Figure 6 System accuracy versus training set size corresponding to the 0.125-resolution dataset (see online version for colours)

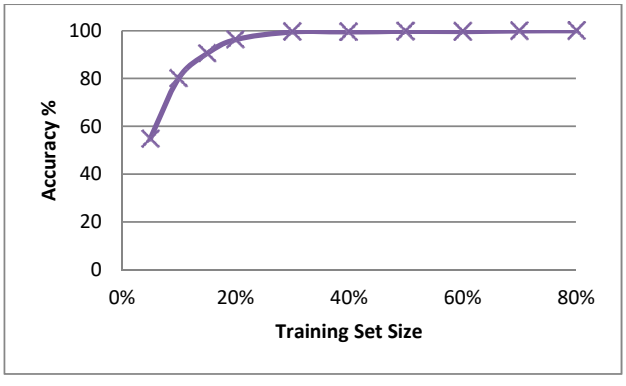


Figure 7 System accuracy versus training set size corresponding to the 0.1-resolution dataset (see online version for colours)

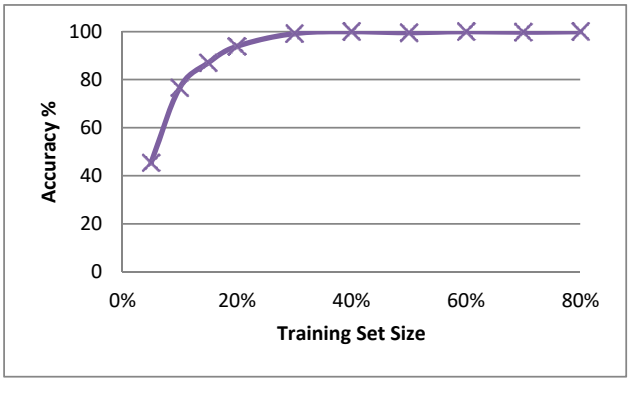


Figure 8 System accuracy versus the number of epochs for the original dataset (see online version for colours)

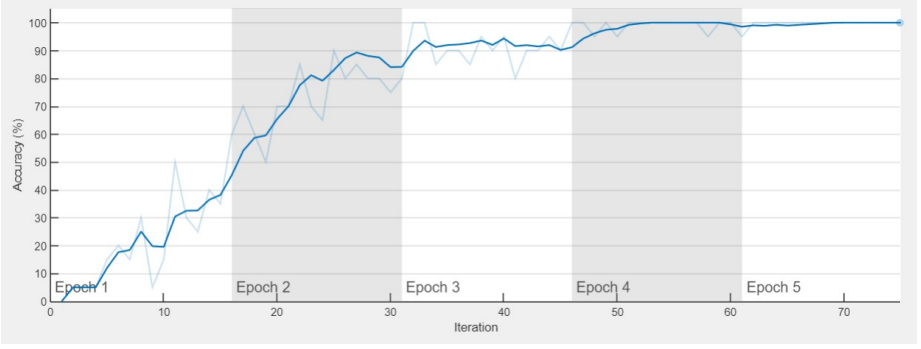


Figure 9 System loss versus the number of epochs for the original dataset (see online version for colours)

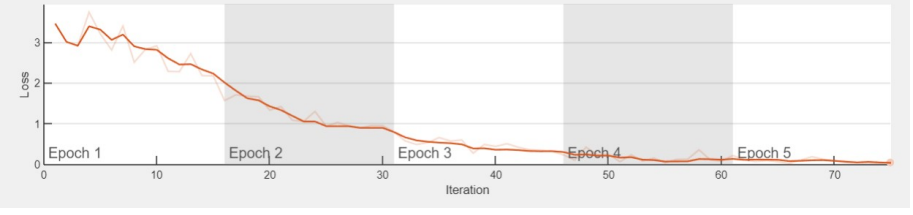


Figure 10 System accuracy versus the number of epochs for the 0.5-resolution dataset (see online version for colours)

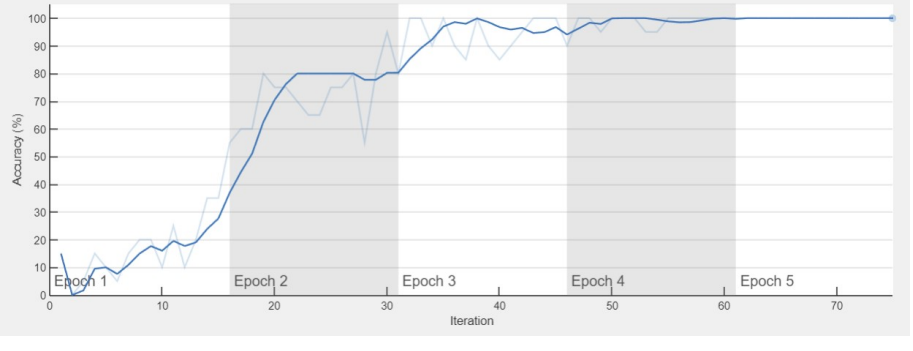


Figure 11 System loss versus the number of epochs for the 0.5-resolution dataset (see online version for colours)

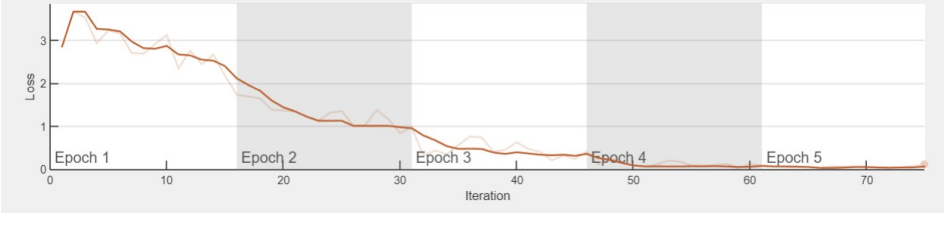


Figure 12 System accuracy versus the number of epochs for the 0.25-resolution dataset (see online version for colours)

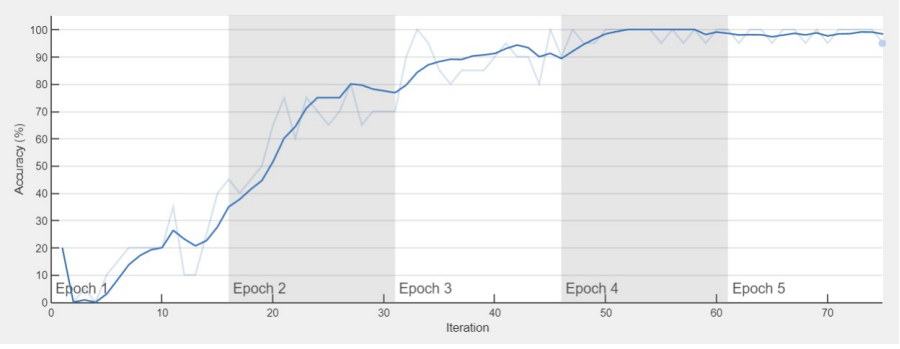


Figure 13 System loss versus the number of epochs for the 0.25-resolution dataset (see online version for colours)

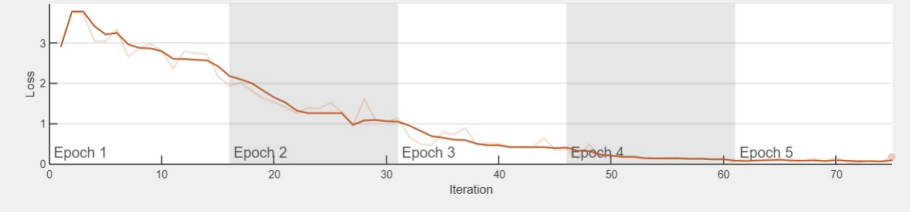


Figure 14 System accuracy versus the number of epochs for the 0.125-resolution dataset (see online version for colours)

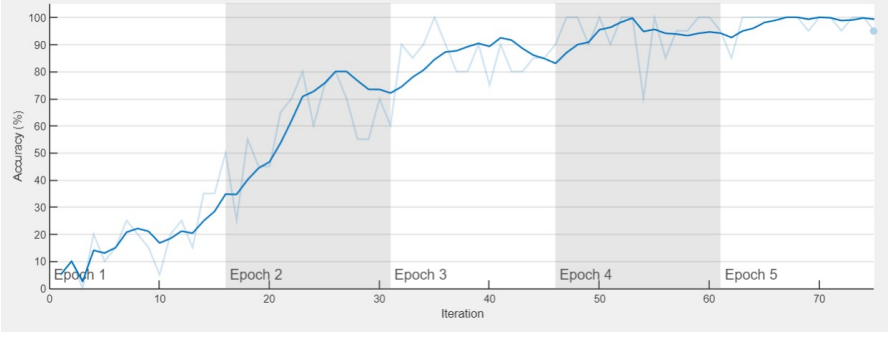


Figure 15 System loss versus the number of epochs for the 0.125-resolution dataset (see online version for colours)

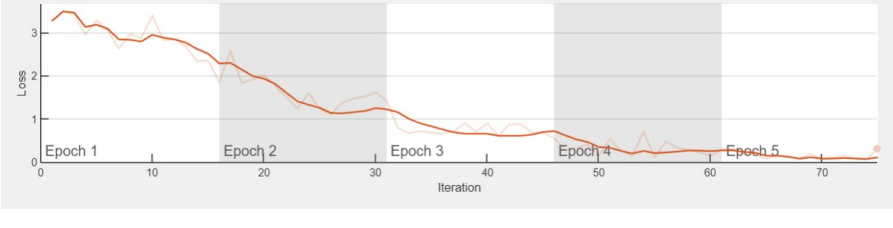


Figure 16 System accuracy versus the number of epochs for the 0.1-resolution dataset (see online version for colours)

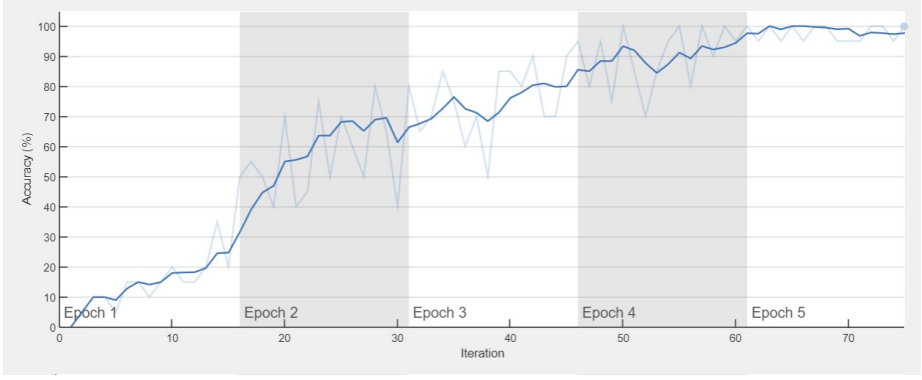


Figure 17 System loss versus the number of epochs for the 0.1-resolution dataset (see online version for colours)

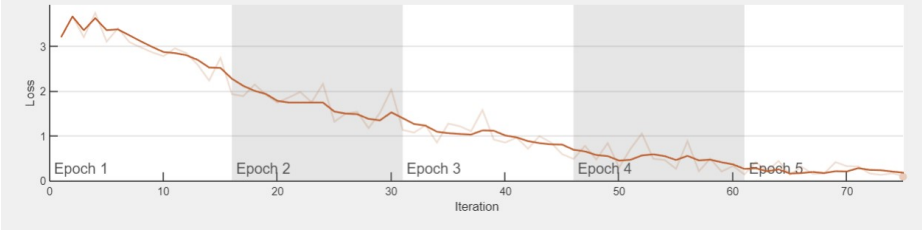


Figure 18 The confusion matrix corresponding to the original dataset (see online version for colours)

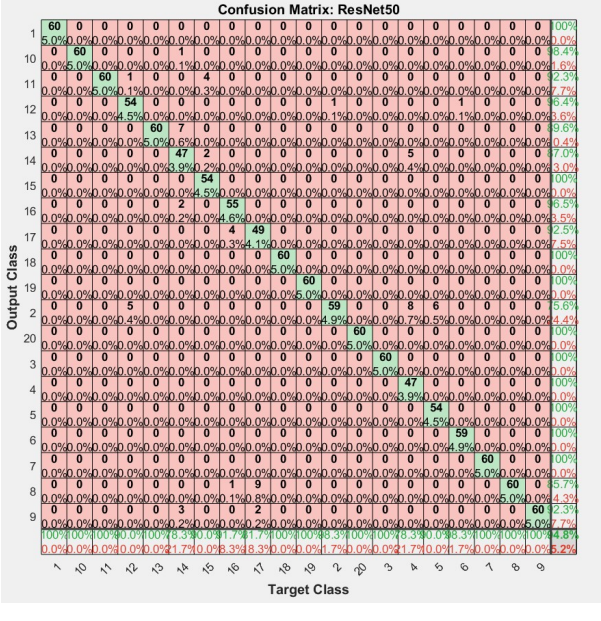


Figure 19 The confusion matrix corresponding to the 0.5-resolution dataset (see online version for colours)

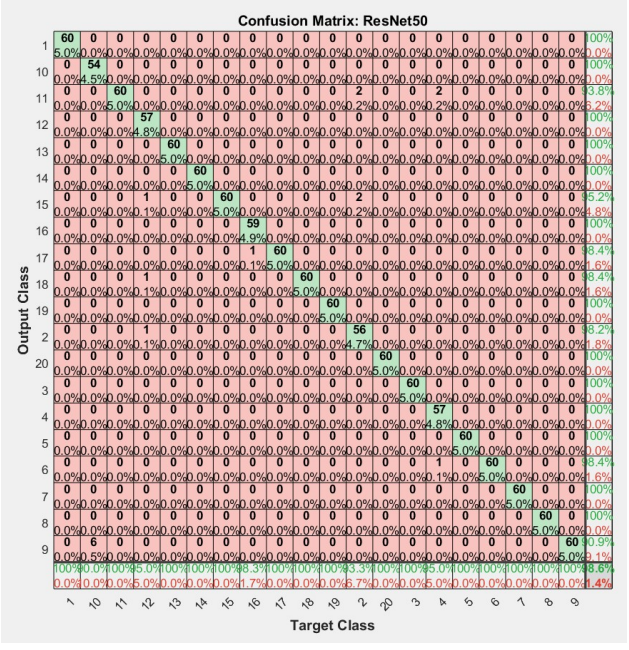


Figure 20 The confusion matrix corresponding to the 0.25-resolution dataset (see online version for colours)

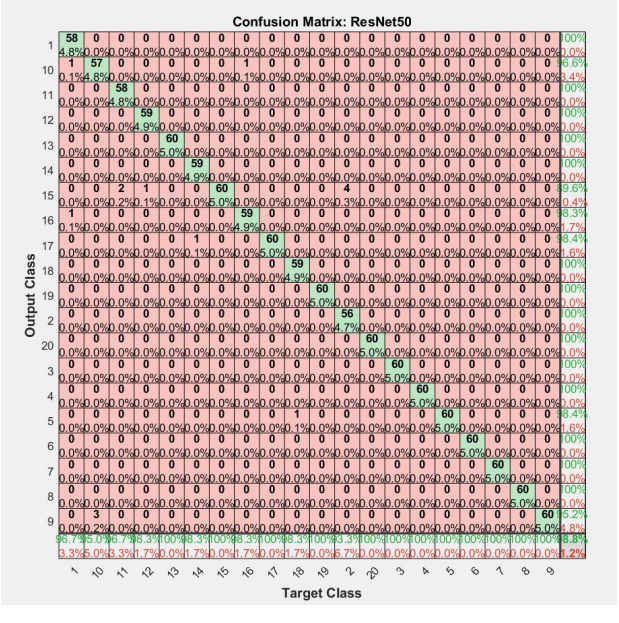


Figure 21 The confusion matrix corresponding to the 0.125-resolution dataset (see online version for colours)

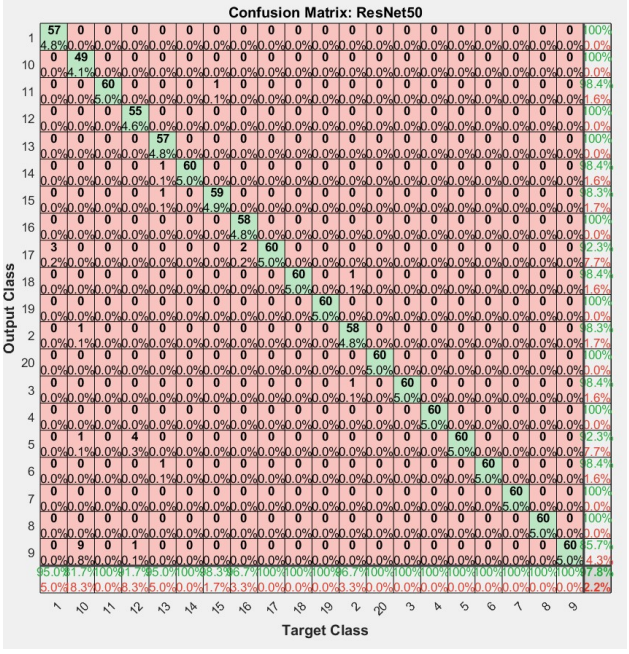


Figure 22 The confusion matrix corresponding to the 0.1-resolution dataset (see online version for colours)

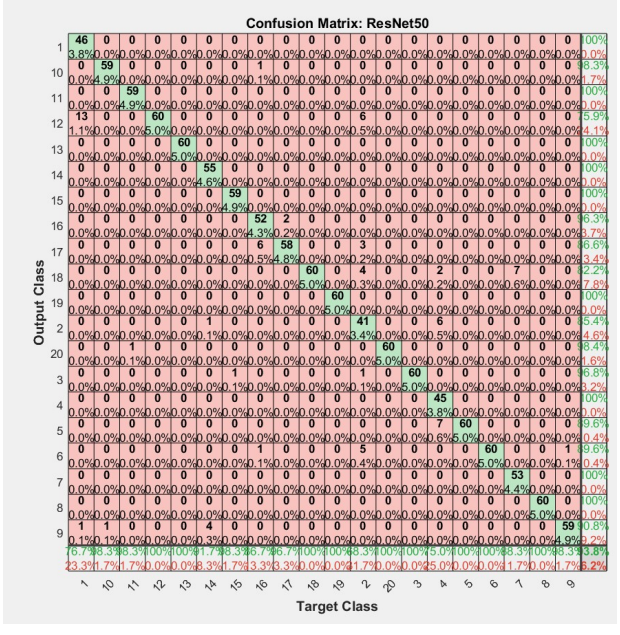


Figure 23 System recall vs. training set size (see online version for colours)

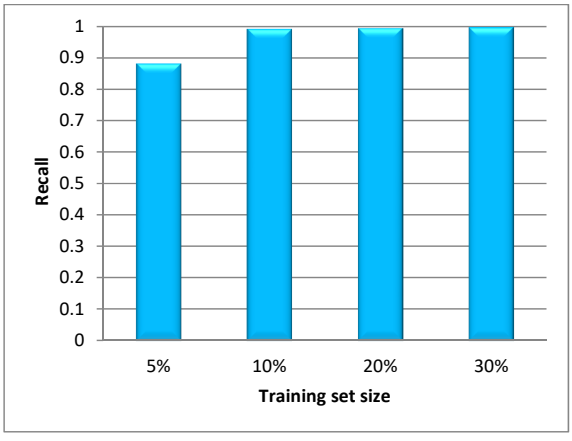


Figure 24 System specificity vs. training set size (see online version for colours)

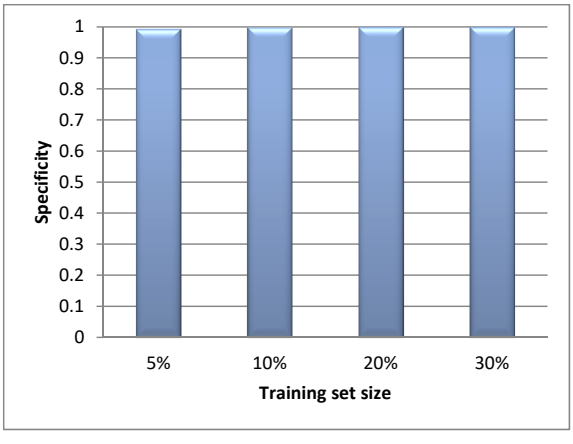


Figure 25 System precision vs. training set size (see online version for colours)

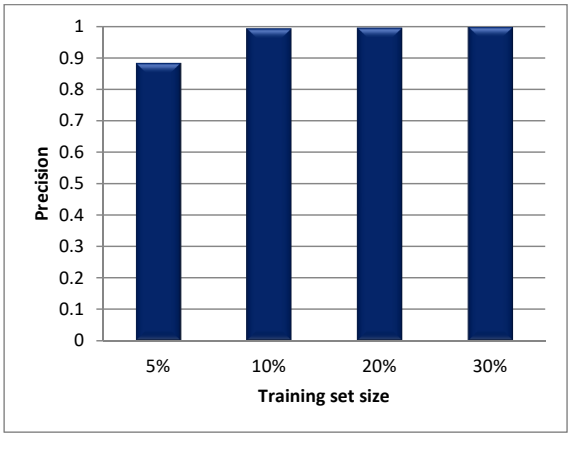


Figure 26 System F-score vs. training set size

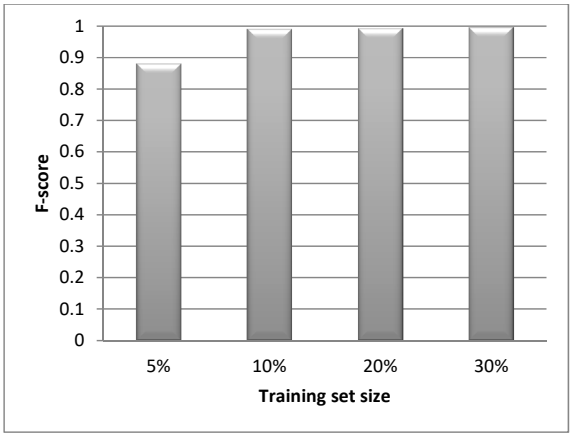


Table 2 Baseline systems' performance based on accuracy rate

Publication	*Year*	*Best accuracy (%)*
Gong et al. (2017)	2017	85.6
Bhowmik et al. (2019)	2019	97.4
Zaeri (2020)	2020	95
Lin et al. (2021)	2021	96.7
Assiri and Hossain (2023)	2023	93.9
Proposed method	2024	99.6

5 Conclusions

Face recognition utilising IR technology can surpass the drawbacks of visible light systems since it is insensitive to changes in facial skin and expression. The anatomical data that IR technology can obtain include underlying traits that are unique to each individual. In this work, we explored CNNs as an emerging area for machine vision and developed a robust system to identify faces in degraded thermal images. The proposed method that implements ResNet-50 architecture can deal efficiently with thermal images that suffer degradation in various ways, including noise, reduced spatial resolution, and temperature drift, in addition to being affected by pose and facial expression changes. The system's performance was evaluated using different measures, including accuracy rate, amount of loss, confusion matrix, recall, specificity, precision, and F-score. The experimental results show that the system has a stable performance and is robust when handling thermal images of low quality resulting from several origins.

References

Assiri, B. and Hossain, M.A. (2023) 'Face emotion recognition based on infrared thermal imagery by applying machine learning and parallelism', *Math Biosci Eng.*, Vol. 20, No. 1, pp.913–929.

Bhowmik, M.K., Saha, P., Singha, A., Bhattacharjee, D. and Dutta, P. (2019) 'Enhancement of robustness of face recognition system through reduced Gaussianity in log-ICA', *Expert Syst. Appl.*, February, Vol. 116, pp.96–107.

Bi, Y., Lv, M., Wei, Y., Guan, N. and Yi, W. (2016) 'Multi-feature fusion for thermal face recognition', *Infrared Physics & Technology*, Vol. 77, pp.366–374.

Gong, D., Li, Z., Huang, W., Li, X. and Tao, D. (2017) 'Heterogeneous face recognition: a common encoding feature discriminant approach', *IEEE Trans. Image Process.*, May, Vol. 26, No. 5, p.20792089.

Gonzalez, R.C. (2018) 'Deep convolutional neural networks [Lecture Notes]', *IEEE Signal Processing Magazine*, Vol. 35, No. 6, pp.79–87.

He, K., Zhang, X., Ren, S. and Sun, J. (2016) 'Deep residual learning for image recognition', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp.770–778.

Hermosilla, G., Rojas, M., Mendoza, J., Farias, G., Pizarro, F.T., San Martin, C. and Vera, E. (2018) 'Particle swarm optimization for the fusion of thermal and visible descriptors in face recognition systems', *IEEE Access*, Vol. 6, pp.42800–42811.

Hermosilla, G., Tapia, D.I.H., Allende-Cid, H., Castro, G.F. and Vera, E. (2021) 'Thermal face generation using StyleGAN', *IEEE Access*, Vol. 9, pp.80511–80523.

Iranmanesh, S., Dabouei, A., Kazemi, H. and Nasrabadi, N.M. (2018) 'Deep cross polarimetric thermal-to-visible face recognition', *IEEE International Conference on Biometrics*, Australia, INSPEC Accession Number: 17934665, DOI: 10.1109/ICB2018.2018.00034.

Kakarwal, S.N., Chaudhari, K.P., Deshmukh, R.R. and Patil, R.B. (2020) 'Thermal face recognition using artificial neural network', *2020 International Conference on Smart Innovations in Design, Environment, Management, Planning and Computing*, pp.300–304.

Kumar, S. and Singh, S.K. (2020) 'Occluded thermal face recognition using bag of CNN', in *IEEE Signal Processing Letters*, Vol. 27, pp.975–979, DOI: 10.1109/LSP.2020.2996429.

Lezama, J., Qiu, Q. and Sapiro, G. (2017) 'Not afraid of the dark: NIR-VIS face recognition via cross-spectral hallucination and low-rank embedding', *IEEE Conference on Computer Vision and Pattern Recognition*, pp.6807–6816.

Lin, C.H., Wang, Z-H. and Jong, G-J. (2020) 'A de-identification face recognition using extracted thermal features based on deep learning', in *IEEE Sensors Journal*, Vol. 20, No. 16, pp.9510–9517, 15 August, DOI: 10.1109/JSEN.2020.2986098.

Lin, S.D. and Chen, K. (2019) 'Thermal face recognition under disguised conditions', *2019 International Conference on Machine Learning and Cybernetics (ICMLC)*, pp.1–7, DOI: 10.1109/ICMLC48188.2019.8949194.

Lin, S.D., Chen, K. and Chen, W. (2019) 'Thermal face recognition based on physiological information', *2019 IEEE International Conference on Image Processing (ICIP)*, pp.3497–3501, DOI: 10.1109/ICIP.2019.8803688.

Lin, S.D., Chen, L. and Chen, W. (2021) 'Thermal face recognition under different conditions', *BMC Bioinformatics*, Vol. 22, No. 5, pp.1–17.

Maeng, H., Liao, S., Kang, D., Lee, S-W. and Jain, A.K. (2012) 'Nighttime face recognition at long distance: cross-distance and cross-spectral matching', *Proc. Asian Conference on Computer Vision*, Korea, pp.1–14.

Pini, S., Borghi, G., Vezzani, R., Maltoni, D. and Cucchiara, R. (2021) 'A systematic comparison of depth map representations for face recognition', *Sensors*, Vol. 21, No. 3, p.944.

Ruuska, S., Hämäläinen, W., Kajava, S., Mughal, M., Matilainen, P. and Mononen, J. (2018) 'Evaluation of the confusion matrix method in the validation of an automated system for measuring feeding behaviour of cattle', *Behavioural Processes*, Vol. 148, pp.56–62.

Silverthorn, D.U. (2015) *Human Physiology*, Jones & Bartlett Publishers, Burlington, MA, USA.

Vigneau, G.H., Verdugo, J.L. and Castro, G.F. (2017) 'Thermal face recognition under temporal variation conditions', *IEEE Access*, Vol. 5, pp.9663–9672, DOI: 10.1109/ACCESS.2017.2704296.

Wu, J. (2017) *Introduction to Convolutional Neural Networks*, LAMDA Group, National Key Lab for Novel Software Technology, Nanjing University, China.

Zaeri, N. and Qasim, R. (2023) 'Thermal image identification against pose and expression variations using deep learning', *Journal of Engineering Research*, ISSN: 2307-1877, available online 4 November 2023, In Press, https://doi.org/10.1016/j.jer.2023.10.043 [online] https://www.sciencedirect.com/science/article/pii/S2307187723003048 (accessed 15 May 2024).

Zaeri, N. (2020) 'Thermal face recognition under spatial variation conditions', *Pattern Recognition and Image Analysis*, Vol. 30, pp.108–124, DOI: 10.1134/S1054661820010174.