Sensors and Materials, Vol. 35, No. 7 (2023) 2567–2579
MYU Tokyo

S & M 3341

Smart Home Surveillance System Based on the Optimized EfficientDet Network

Ming-Tsung Yeh,[1*] Yu-Chi Tsai,[2] Chi-Huan Cheng,[2]
Yi-Nung Chung,[2] and Pei-Syuan Lu[2]

[1]Department of Electrical Engineering, National Chin-Yi University of Technology,
57, Sec. 2, Zhongshan Rd., Taiping Dist, Taichung 411030, Taiwan
[2]Department of Electrical Engineering, National Changhua University of Education,
No. 1, Jinde Rd., Changhua City, Changhua County 50007, Taiwan

(Received December 30, 2023; accepted June 14, 2023)

Keywords: smart surveillance system, full-time surveillance, face recognition, EfficientDet network, auto-coloring

Home security systems have been extensively used to protect our property and family, and these products are always equipped with some devices, including indoor and outdoor cameras, infrared motion sensors, human body temperature sensors, and smart locks. Security systems are designed to detect the presence of people or moving objects. However, they have certain limitations. Firstly, these systems are inactive when not turned on, rendering them ineffective during those times. Additionally, false alarms are common during system surveillance, posing a challenge to the system's overall reliability. A smart surveillance camera has recently been added to the system, but it cannot distinguish between family members and intruders. A full-time facial recognition system has been proposed in this paper to address the drawbacks of the current security system. The Day Night Surveillance Neural Network (DNSNN), which is a face recognition network based on the optimized EfficientDet, is proposed. The DNSNN provides full-time recognition in this study and divides the system into day and night modes. It uses visible light images in day mode under good light conditions to perceive objects. Under poor light conditions, the camera automatically takes grayscale images with near IR. However, in these images, objects are challenging to recognize, and thus the accuracy rate is reduced. A proposed auto-coloring system is applied to colorize the grayscale images. The colorized images can have a similar hue to the visible light images and are equipped with the same vision. This method can improve system recognition capabilities under dim light. The experimental results show that our proposed approaches recognize family members and intruders under all light conditions and have an identification accuracy of more than 90%. This system can achieve full-time smart home surveillance.

*Corresponding author: e-mail: mtyeh@ncut.edu.tw
https://doi.org/10.18494/SAM4297

ISSN 0914-4935
https://myukk.org/

1. Introduction

Home safety encompasses fall prevention, monitoring carbon monoxide levels, and crime prevention, all of which are crucial aspects of society. Crime prevention explicitly involves home surveillance. Unfortunately, many households lack a surveillance system, making it challenging to apprehend criminals. To mitigate the risks associated with life and property loss, people have become increasingly concerned with implementing home surveillance, as criminals often break into homes to steal, rob, or kidnap individuals. In earlier years, hiring security personnel such as building guards or community security was common to address these concerns. However, as users recognized the potential of home surveillance systems to enhance security while reducing the need for human resources, the installation of surveillance systems gained popularity. Integrating home environments and the Internet of Things (IoT) has given rise to the flourishing concept of smart homes in recent years.(1) Smart homes can detect various environmental factors such as temperature, humidity, and carbon monoxide levels. Users have the ability to remotely control appliances and receive real-time data to maintain an optimal home environment. Within smart homes, surveillance systems employ cameras to monitor home conditions and provide feedback to users, enabling them to manage their smart homes remotely. Artificial intelligence (AI) has introduced significant convenience by offering features such as face recognition,(2) speech recognition, and intelligent responses.(3) The traditional surveillance system relies on sensors, cameras, and message transmission. However, it requires constant human monitoring, resulting in significant inconveniences. To achieve complete automation and reduce the reliance on human resources, a smart home combines cameras with face recognition to create a new surveillance system. This system only requires a camera at the door, where face recognition technology can identify intruders and authorized members. If an intruder is detected, the system can alert the members. Integrating cameras and an Artificial Intelligence of Things (AIOT) architecture enhances home security. However, this approach still has limitations.(4) Conventional cameras struggle to recognize objects under low-light conditions, leading to decreased accuracy. During the nighttime, thermal infrared cameras are the only viable option, but they can only generate grayscale images and cannot perform recognition on the captured images. Criminals often use low-light environments to carry out illegal activities, making the nighttime critical for home security. If we fail to address the challenges posed by nighttime environments, the need for human security personnel to monitor the surroundings will persist. This presents a significant challenge for automated surveillance systems.

Object detection has witnessed significant advancements in deep learning, improving accuracy and efficiency. This field can be divided into two modes.(5) First, early object detection algorithms commonly use a two-stage detector, such as a Region-based Convolutional Neural Network (R-CNN) and Fast R-CNN.(6,7) Second, the one-stage detector, employed in YOLO and EfficientDet, has become mainstream owing to its superior efficiency compared with the two-stage detector.(8,9) In the two-stage detector, features are extracted to generate a feature map, which is then used for object localization. Subsequently, image classification is performed. On the other hand, algorithms in the one-stage detector are more efficient because they simultaneously calculate object localization and classification. Object detection algorithms find applications in various domains, including traffic safety, defect detection, and face recognition.

In recent years, object detection has been extensively utilized in face recognition. For instance, Jiang and Learned-Miller applied the Faster R-CNN model to face recognition, enhancing its accuracy and speed using the WIDER dataset.[10] Yang and Jiachun employed YOLO for face recognition and observed that YOLOv3 offered shorter detection periods and greater robustness.[11] Awais *et al.* integrated face recognition into a surveillance system, comparing faces captured by the camera with those in the database and alerting the user if a face is not recognized.[12] These methods have demonstrated sufficient speed and detection precision. It is worth noting that face recognition algorithms typically require well-lit conditions during the day to capture the entire face. Image restoration techniques are employed during nighttime to match daylight conditions and improve accuracy.

Using light compensation theory, Bao and Dang improved preprocessing techniques, resulting in an enhanced MCTNN model.[13] Experimental results demonstrated that MCTNN achieved increased accuracy in low-luminosity environments. Liang *et al.* proposed the REGDet face detection architecture designed to detect faces under low-luminosity conditions.[14] These methods effectively address the challenges of low luminosity and unrecognizable faces but do not extend to completely luminosity-free environments. In such environments, only thermal infrared cameras can be used, which capture grayscale images, whereas visible light cameras capture RGB images. Grayscale images need to be colorized to apply face recognition in these scenarios. Wu *et al.* improved the Generative Adversarial Network (GAN) and introduced a colorization architecture for grayscale images, yielding superior results in their experiments.[15] Li *et al.* utilized broad-GAN to achieve remarkable colorization results for grayscale images while enhancing the training stability of GAN.[16] Ji *et al.* proposed MC-GAN for coloring Synthetic Aperture Radar (SAR) images and demonstrated its superiority over other methods in coloring SAR images.[17]

In our system, the Day Night Surveillance Neural Network (DNSNN) is proposed to enhance recognition efficiency and improve the recognition effect for full-time smart home surveillance. The key innovation lies in the division of day and night modes. The surveillance system must consider both the speed and accuracy of face recognition. EfficientDet, an algorithm with high accuracy and computational efficiency, is suitable for smart homes as it performs at par with other algorithms. Consequently, many researchers have adopted the lightweight EfficientDet as the network for recognition-based tasks. Additionally, we have incorporated the proposed auto-coloring system to convert grayscale images to RGB images, addressing the issue of poor recognition during nighttime. The coloring process is accomplished using a combination of GAN and autoencoder. The resulting colored images are fed into the neural network for member or intruder recognition. In this study, the EfficientDet algorithm is integrated with the auto-coloring system, enabling the smart home identification system to achieve improved accuracy during the day and night.

2. Proposed Method

The overall system architecture is depicted in Fig. 1. The system utilizes a camera integrated with deep learning techniques to enable surveillance capabilities during nighttime. The camera captures images of family members, which are then stored in a database. The database undergoes

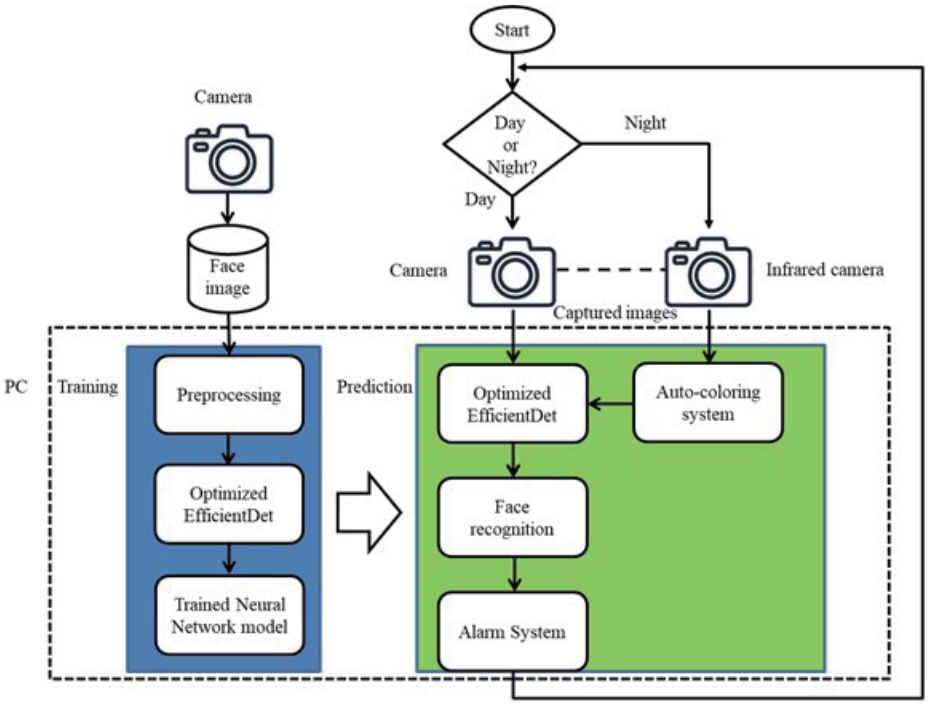

Fig. 1. (Color online) System architecture.

data augmentation, and deep learning algorithms are employed to train a model specifically for family member recognition. The surveillance system for family members determines whether it is operating in the day or night mode. During the day mode, a visible camera captures images of both members and potential intruders. Face recognition algorithms are applied to determine the identity of individuals as either members or intruders. The system alerts the user if an individual is identified as an intruder. An infrared camera is utilized in the night mode, resulting in grayscale images. To overcome this limitation, an auto-coloring system adds color to the images before the face recognition algorithms process them.

2.1 Automatic face labeling application program

In this study, we utilized a Logitech camera to capture facial images of family members, which were then used as a database. In conventional object detection methods, much time is typically spent on face labeling. However, in this study, we have developed a face labeling application programming interface (API) that simplifies this process. In the API, a square is drawn on the screen, as illustrated in Fig. 2. When the camera captures an image, we align the square with the face in the image, allowing us to create face images and training sets automatically. The training sets are stored as XML files containing the coordinates of the

square, namely, xmin, ymin, xmax, and ymax. These images, generated using the four coordinates, are fed into the object detection system for further processing.

Data augmentation plays a crucial role in increasing the accuracy of training sets. It involves applying various image processing techniques to enhance the dataset. This study uses salt and pepper noise, blur, and Gaussian noise as augmentation techniques. Salt and pepper noise refers to introducing white and black pixels resembling salt and pepper sprinkled on the image. The blur effect is achieved by applying a low-pass filter through convolution, resulting in blurred vision. The surveillance system is designed to recognize both blurry and high-resolution images effectively. Gaussian noise, a commonly used technique in data augmentation, is also incorporated. However, we do not apply zooming, scaling, or rotation transformations owing to the face labeling process, which requires precise coordination. Using such transformations can cause the face to move out of the specified coordinates, leading to the issue of vanishing gradients. The data augmentation process is illustrated in Fig. 3.

2.2 Face recognition

EfficientDet, proposed by the Google team in 2020, is employed as a neural network for face recognition.[9] This algorithm demonstrates a faster identification than existing algorithms. The EfficientDet architecture consists of several backbone networks, namely, EfficientNet-B0 to B6. These backbone networks are combined with a bidirectional feature pyramid network (BiFPN)

Fig. 2. (Color online) Face labeling application programming interface.

Fig. 3. (Color online) Data augmentation.

to form the feature network. Finally, the network is connected to the class and box prediction networks. The architecture of EfficientDet is illustrated in Fig. 4.

BiFPN is a highly accurate feature network component created by combining Cross-Scale Connections and Weighted Feature Fusion. Cross-Scale Connections involve reducing two nodes and connecting the input and output nodes in PANet. Weighted Feature Fusion compares three fusion methods: unbounded fusion, softmax-based fusion, and fast normalized fusion. Unbounded fusion has the disadvantage of unbounded scalar weights, which can result in unstable training. Softmax-based fusion applies the softmax function to the weights, restricting them to a range between 0 and 1. However, the use of multiple softmax operations can increase computation time. The fast normalized fusion formula Eq. (1) is employed to reduce latency. In Eq. (1), the weights (*wi*) are ensured to be greater than or equal to 0 through the application of ReLU. Additionally, a small value (ε = 0.0001) is added to avoid numerical instability. In this study, the softmax operation is not applied to reduce the computational load on the GPU, resulting in improved efficiency.

$$O = \sum_i \frac{w_i}{\varepsilon + \sum_j w_j} . I_i \tag{1}$$

Here, O is output, w is weight, ε is a small value set to 0.0001, and I is input.

2.3 Auto-coloring system

We convert our infrared images' luminance (L) channel into visible light images, L channel in the Lab color space. This conversion process is achieved by combining the training results of GAN, where the "ab" channel is used as input for coloring. Furthermore, the auto-coloring mechanism is implemented. The first stage of the system architecture is depicted in Fig. 5. It involves transferring the infrared image to the Lab color space and extracting the L channel. A batch normalization layer is added to every stack of layers to enhance image diversity and

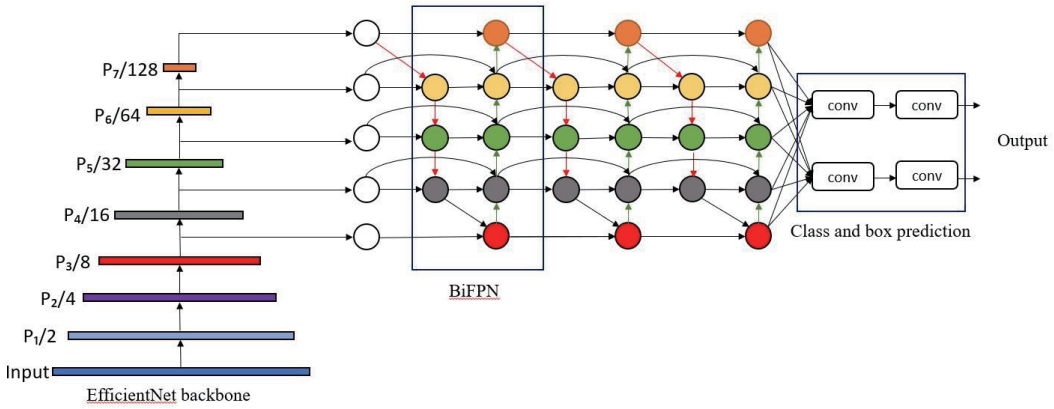

Fig. 4. (Color online) EfficientDet framework.

adaptability. Additionally, residual layers are incorporated as connections to efficiently extract differences between two L channel feature maps, facilitating the network's learning process.

The second stage of the system architecture involves the use of visible light L images as input. The goal is to predict the ab color space for these visible light images. To achieve this, we improve the optimization and generator of GAN. We also modify the loss function to apply the eLU activation function, which ensures that the values do not become negative during the RGB to CIELAB color space conversion. Additionally, we add color elements to the infrared images to align them with the visible light images. This part of the architecture also involves model training, similar to the GAN model.

The discriminator is trained through the GAN framework and affects the generator based on input and output. The objective is to generate ab channels using GAN. The architecture diagram is presented in Fig. 6, where the input is a (256, 256, 1) image. This work employs a series of convolution and deconvolution modules with varying kernels, resulting in six modules. The generator output is (256, 256, 2), which serves as the input for the discriminator (256, 256, 2). The discriminator shares the same structure as the generator, and its output generates two types of images. The first type is reconstructive, where the output of the generator is used as the input for the discriminator. The generated results pass through the network and produce reconstructed images. The second type is retrue, where the discriminator generates the input ab channels of the visible image. These two outputs are calculated using the mean absolute error (MAE). The

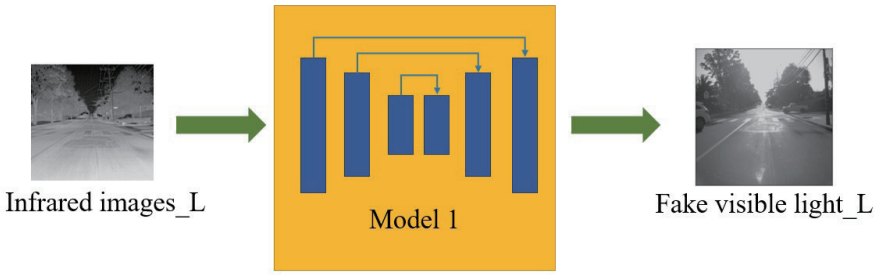

Fig. 5. (Color online) First stage of auto-coloring system architecture.

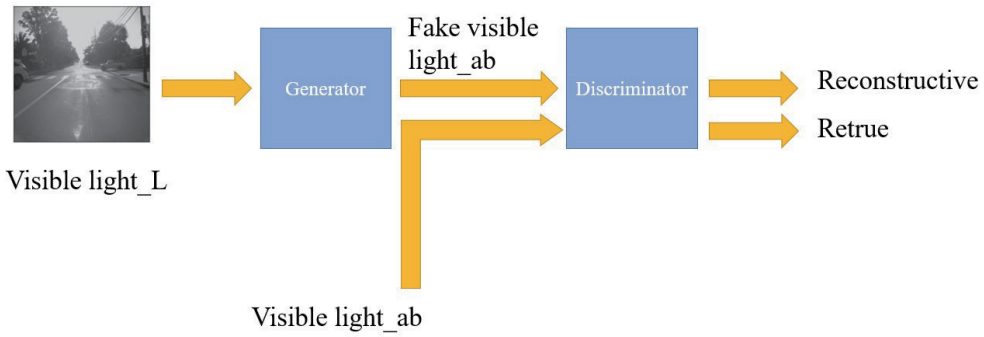

Fig. 6. (Color online) Second stage of auto-coloring system architecture.

discriminator is corrected by adjusting its parameters based on the MAE results. The MAE also serves as the loss function. Consequently, this loss function corrects the generator. After 20000 epochs, the distribution of the ab channels is more aligned with the expected results.

2.4 DNSNN

The architecture of DNSNN combines the automatic generation of training data, EfficientDet, and the auto-coloring system. The flowchart of the training process is depicted in Fig. 7. It starts with the camera capturing images of family members, which are then processed by an automatic face labeling system to generate the training data. This training data is used as input for EfficientDet. The EfficientDet model is trained using the generated training data, and the trained model is exported and saved for further use.

The DNSNN architecture consists of two modes: day and night. The day mode of the surveillance system is depicted in Fig. 8. In this mode, a visible light camera captures facial images for recognition purposes. The system determines whether the person is a member or an intruder. If the person is recognized as a member, the system does not send any reminders to the user. However, if the person is identified as an intruder, the system sends a reminder to the user. The architecture of the night mode is illustrated in Fig. 9. In the night mode, the system utilizes

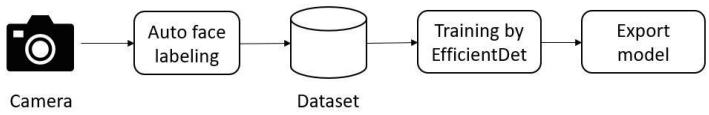

Fig. 7. Flow chart of training.

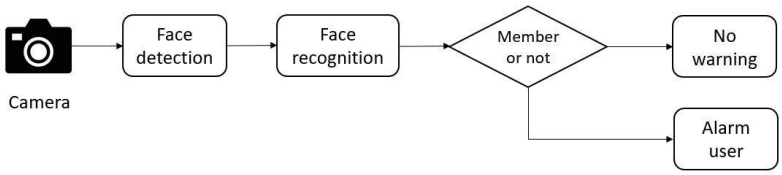

Fig. 8. Architecture of day mode.

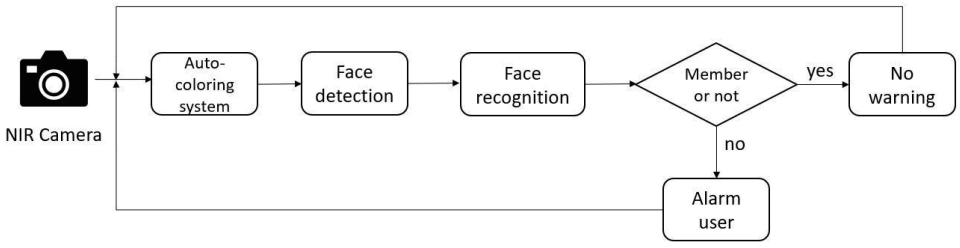

Fig. 9. Architecture of night mode.

Table 1.
Confusion matrix.

	Ground true positive	Ground true negative
Predict positive	*TP*	*FP*
Predict negative	*FN*	*FN*

an infrared camera to capture grayscale images. Since the grayscale images lack color information, they must be colorized using a neural network. Once the grayscale image is colorized, it is sent to the face recognition module for identification.

3. Experimental Results and Discussion

In this paper, we introduced DNSNN, a novel smart home surveillance system. The system incorporates an automatic face labeling interface that generates efficient training data, thereby facilitating the addition of new members. By leveraging the EfficientDet algorithm, our proposed network achieves shorter processing times in training and prediction, and enables real-time alerts for users. Additionally, the auto-coloring system addresses face recognition challenges in low-light environments, enabling accurate recognition comparable to daytime performance.

In the experiments, we evaluated the performance of each deep learning model. The evaluation process involved four parameters: true positive (*TP*), true negative (*TN*), false positive (*FP*), and false negative (*FN*). The results were analyzed using a confusion matrix, as shown in Table 1. *Accuracy* [Eq. (2)], Precision [Eq. (3)], and *Recall* [Eq. (4)] were computed to assess the models.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{2}$$

$$Precision = \frac{TP}{TP + FP} \tag{3}$$

$$Recall = \frac{TP}{TP + FN} \tag{4}$$

The Intersection over Union (*IOU*) is a standard metric used in object detection to measure the overlap between the predicted bounding box and the ground truth. It quantifies the accuracy of the answer by calculating the area of overlap between two objects in the model. If the *IOU* is greater than a predefined threshold (typically 0.5), the target is considered a true positive (*TP*). The target is considered a false positive (*FP*) if the *IOU* is smaller than the threshold. *Precision* and *Recall* are calculated from these object detection results, and the corresponding diagram can be used to illustrate them. The evaluation process involves computing the average precision (*AP*), which is the sum of the area under the *Precision–Recall* curve. The mean average precision (*mAP*) is then calculated as the average *AP* across different types of object recognition.

In this study, we collected a dataset consisting of 500 images and three classes, namely, "member", "member1", and "intruder". The "member" and "member1" classes represent two family members, while the "intruder" class contains images of other people. Initially, the results could have been more satisfactory, but we collected more intruder data to improve performance. The augmentation technique proposed by the automatic face labeling application programming interface significantly increased the training dataset size to 3500 images, resulting in a *mAP* of 0.9. This demonstrates that the accuracy improves rapidly with the use of the automatic face labeling application programming interface. A comparison of the results obtained using two different training approaches is presented in Table 2.

In the nighttime experiment, the face recognition process initially relies on the day mode prediction. A NIR camera is used to capture photos with a grayscale level, but the prediction results have low accuracy. The grayscale images are shown in Fig. 10. To improve the recognition accuracy, an auto-coloring system is applied to the grayscale images. The colorized images are depicted in Fig. 11. After undergoing auto-coloring, the colorized images are fed into the face recognition process, resulting in predictions with the same level of accuracy as the day mode. It is observed that directly using grayscale images for prediction yields lower accuracy than using the images after applying the proposed auto-coloring system. A comparison of the accuracies is presented in Table 3. The detected results are illustrated in Fig. 12 for predictions from the colorized images.

Table 2.
Comparison of results obtained by two types of training.

	mAP	AP50
Without AFLAPI	0.75	0.92
AFLAPI	0.94	0.97

Fig. 10. Grayscale images captured by NIR camera.

Fig. 11. (Color online) Grayscale images are colorized by the proposed auto-coloring system.

Table 3 Comparison of accuracies.

	mAP	AP50
Night mode	0.92	0.96
Day mode	0.94	0.97
Grayscale	0.41	0.61

Fig. 12. (Color online) Predictions obtained from the colorized images.

4. Conclusions

In this paper, we introduced DNSNN, a smart home surveillance system designed for recognition purposes. The system incorporates an automatic face labeling application program that generates efficient training data, thereby facilitating the identification of family members. By adopting the EfficientDet algorithm, object recognition is performed more swiftly, reducing computational time and achieving faster recognition results. Additionally, the auto-coloring system addresses the challenges of face recognition in low-light environments. Remarkably, the proposed approaches demonstrate high accuracy in both daytime and nighttime face recognition scenarios, surpassing a 90% identification accuracy threshold. The experimental results validate the system's ability to recognize family members and intruders under various lighting conditions, establishing its potential for comprehensive and continuous smart home surveillance.

Acknowledgments

This work was supported by the National Science and Technology Council under Grant no NSTC 111-2222-E-167-003.

References

1 K. Balasubramanian and A. Cellatoglu: IEEE Trans. Consumer Electronics **54** (2008) 1681. https://doi.org/10.1109/TCE.2008.4711220
2 W. Wu, Y. Yin, X. Wang, and D. Xu: IEEE Trans. Cybernetics **49** (2019) 4017. https://doi.org/10.1109/TCYB.2018.2859482
3 J. W. Kim, H. Yoon, and H. Y. Jung: Proc. IEEE Access (IEEE, 2021) 136476–136486. https://doi.org/10.1109/ACCESS.2021.3115608
4 E. Baccour, N. Mhaisen, A. A. Abdellatif, A. Erbad, A. Mohamed, M. Hamdi, and M. Guizani: IEEE Commun. Surv. Tutorials 24 (IEEE, 2022) 2366. https://doi.org/10.1109/COMST.2022.3200740

5 X. Yang, X. Zhang, N. Wang, and X. Gao: IEEE Trans. Geoscience and Remote Sensing **60** (2022) 1. https://doi.org/10.1109/TGRS.2021.3128060
6 R. Girshick, J. Donahue, T. Darrell, and J. Malik: Proc. 2014 IEEE Conf. Computer Vision and Pattern Recognition (CVPR, Columbus, 2014) 580–587. https://doi.org/10.1109/CVPR.2014.81
7 R. Girshick: Proc. 2015 IEEE Int. Conf. Computer Vision (2015) 1440–1448. https://doi.org/10.1109/ICCV.2015.169
8 J. Redmon, S. Divvala, R. Girshick, and A. Farhadi: Proc. 2016 IEEE Conf. Computer Vision and Pattern Recognition (2016) 779–788. https://doi.org/10.1109/CVPR.2016.91
9 M. Tan, R. Pang, and Q. V. Le: Proc. 2020 IEEE/CVF Conf. Computer Vision and Pattern Recognition (2020) 10778–10787. https://doi.org/10.1109/CVPR42600.2020.01079
10 H. Jiang and E. Learned-Miller: Proc. 2017 12th IEEE Int. Conf. Automatic Face & Gesture Recognition (2017) 650–657. https://doi.org/10.1109/FG.2017.82
11 W. Yang and Z. Jiachun: Proc. IEEE Int. Conf. Knowledge Innovation and Invention (2018) 221–224. https://doi.org/10.1109/ICKII.2018.8569109
12 M. Awais, M. J. Iqbal, I. Ahmad, M. O. Alassafi, R. Alghamdi, M. Basheri, and M. Waqas: Proc. IEEE Access (IEEE, 2019) 121236–121244. https://doi.org/10.1109/ACCESS.2019.2937810
13 Y. Bao and R. Dang: Proc. 2021 2nd Int. Conf. on Artificial Intelligence and Computer Engineering (2021) 704–707. https://doi.org/10.1109/ICAICE54393.2021.00138
14 J. Liang, J. Wang, Y. Quan, T. Chen, J. Liu, H. Ling, and Y. Xu: IEEE Trans. Multimedia **24** (2022) 1609. https://doi.org/10.1109/TMM.2021.3068840
15 W. Wu, Y. Li, and C. Che: Proc. 2021 Int. Conf. on Computer, Blockchain and Financial Development (2021) 45–48. https://doi.org/10.1109/CBFD52659.2021.00016
16 H. Li, B. Sheng, P. Li, R. Ali, and C. L. P. Chen: IEEE Trans. Image Processing **30** (2021) 8526. https://doi.org/10.1109/TIP.2021.3117061
17 G. Ji, Z. Wang, L. Zhou, Y. Xia, S. Zhong, and S. Gong: IEEE Geosci. Remote Sens. Lett. **18** (2021) 296. https://doi.org/10.1109/LGRS.2020.2969891

About the Authors

Ming-Tsung Yeh received his M.S. and Ph.D. degrees from National Changhua University of Education, Taiwan, in 2012 and 2016, respectively. Since 2022, he has been an assistant professor at National Chin-Yi University of Technology, Taiwan. His research interests are in AI, deep learning, image processing, and intelligent control. (mtyeh@ncut.edu.tw)

Yu-Chi Tsai received his B.A. degree from Nanhua University, Taiwan, in 2020. He is currently working toward his M.S. degree in electrical engineering at National Changhua University of Education, Changhua, Taiwan. His research interests are in AI, deep learning, image processing, and intelligent control. (alex861220@gmail.com)

Chi-Huan Cheng received his B.A. degree from Tunghai University, Taiwan, in 2020. He is currently working toward his M.S. degree in electrical engineering at National Changhua University of Education, Changhua, Taiwan. His research interests are in machine learning, image processing, and intelligent control. (webber8844@gmail.com)

Yi-Nung Chung received his Ph.D. degree from Texas Tech University, Lubbock, TX, USA, in 1990. He is currently a professor at National Changhua University of Education, Changhua, Taiwan. He is also the Dean of Academic Affairs. His research interests include image processing and computer vision. (ynchung@cc.ncue.edu.tw)

Pei-Syuan Lu received her M.S. degree in electrical engineering from National Changhua University of Education, Taiwan, in 2022. She is currently working in DELTA ELECTRONICS, INC, Taiwan. Her research interests are in machine learning, image processing, and power electronics. (angellups1999@gmail.com)