Proceedings of the 6th International Conference on Computing and Data Science
DOI: 10.54254/2755-2721/86/20241584

Multi-model fusion of DSFD and Zero-DCE for low-light face detection

Pei Li

College of Computer Science, Xiamen University Malaysia, Selangor, Malaysia

AIT2209076@xmu.edu.my

Abstract. Due to the wide application of face detection, face recognition task under extreme lighting conditions inevitably becomes one of the challenges. Under low-light conditions, the face detection task becomes more complex and difficult due to the presence of more noise and lower visibility in the image. For this extreme environment, this paper proposes a multi-model-assisted solution that combines image enhancement and face detection models to handle this task. The DarkFace dataset was used in this study's validation of a number of distinct face identification models and picture enhancement techniques. The experimental findings demonstrate that, on this dataset, the DSFD (Dual Shot Face Detector) and Zero-DCE (Zero-Reference Deep Curve Estimation) methods perform noticeably better than the other techniques, achieving a recall of 24.61%. Therefore, these two algorithms are selected as the core processing flow in this paper to improve the facial detection system's low-light performance. This combination offers a workable answer to the problem of face detection in low light by effectively increasing the visibility of the picture and raising the accuracy of face detection.

Keywords: low-light conditions, face detection, deep learning, image enhancement

1. Introduction

When we casually pick up our cell phones in the morning, they are then unlocked based on the face captured by the camera. A self-driving tram traveling on the highway recognizes whether the object in front of it is a vehicle or a pedestrian based on its vision system. As object recognition technology becomes more sophisticated, object detection is being used in more applications and more and more devices need to recognize objects. Existing research has combined object detection and object recognition through shared volume features to further recognize the specific class or identity of the object on the basis of the object detection frame to achieve the effect of improving the efficiency and accuracy of the overall system [1]. Among them, the most basic object detection technology will become an important part of the object recognition technology, which is used to realize the accurate recognition of all the objects in the frame by the auxiliary object recognition system. The object recognition and object detection under different lighting conditions will become a great challenge, low light conditions have been proven to interfere with traditional single-stage object detection models such as YOLOv5, resulting in a lack of accuracy [2], which is an important challenge for the current object detection technology, because low light conditions are the most common kind of lighting conditions at present, and especially under extreme low light conditions how to accurately detect the corresponding objects is also an important challenge. This paper present a processing framework in this research that is built on

DOI: 10.54254/2755-2721/86/20241584

pre-existing models, which provides an alternative process for low-light face detection tasks without additional training and building new models, improving the recognition accuracy of traditional models without increasing additional training expenses. It also provides a contribution to the field of multi-model assistance for complex tasks. This thesis focuses on face identification in low light situations. To improve the model's performance, low light photos are first preprocessed using the Zero-DCE model to improve their visibility, and then they are sent through the DSFD model for face detection.

2. Related work

The computer vision technique for localizing items in an image is target detection. This approach is still difficult in low-light situations even if it operates at a human level in typical circumstances [3].

The detection of objects in low light conditions has been the subject of several research projects. By improving low light images, we can partially address this issue. Manual feature methods, such as histogram equalization, adaptive histogram equalization, gamma correction, and decomposition models based on Retinex theory, are one type of low light image enhancement techniques. The other approach is deep learning based algorithms, a reference model with supervised learning using pairs of low and normal lighting graphics and a reference-free model that does not require pairs of data and learns by methods such as GAN. All these methods can significantly enhance the visibility of the image and give a better image input to the target detection model. And according to Kim et al. and comparisons found that deep learning based methods outperform manual feature methods in most cases, especially when the training dataset and test dataset have the same distribution [4].

As for target detection in low-light conditions, there are different aspects of research directions. A part of the research focuses on improving the existing models by changing part of the network of the existing models to a structure that is more suitable for low-light environments, e.g., IDOD-YOLOv7, DK-YOLOv5 and other models focus on improving the existing model frameworks, and DK-YOLOv5 enhances the feature extraction by introducing EnlightenGAN as enhancement algorithm, and by improving the SPPF module to an R- SPPF module to enhance the feature extraction capability and improve the inference speed, which improves the accuracy of DK-YOLOv5's detection in low light [2], whereas IDOD-YOLOv7 constructs a SAIP module that includes four filters, namely white balance, gamma correction, contrast enhancement and image sharpening, and uses a small CNN network for parameter estimation to improve the image enhancement effect. However, they are not without drawbacks, as their demand for computational resources and memory may make real-time detection impossible for resource-poor devices. Another part of the research focuses on multi-feature fusion, i.e., fusing different feature data and low-light RGB images for target detection, and using different feature-assisted models to improve the accuracy.

For example, using depth data or infrared light data to assist in target detection based on the original image. A two-path convolutional neural network was created by Lin et al. to analyze color and depth pictures independently, fusing color and depth features before classification, and maintaining the enhanced features through a join operation, demonstrating excellent accuracy [5]. For a multispectral object detection framework, Thaker et al. suggested an approach based on a Faster R-CNN and feature pyramid network, which enables the model to provide accurate object identification in dimly lit environments by fusing visual and thermal imaging features, and improves the robustness of the model under different light conditions [6]. This type of approach may have high accuracy because it introduces additional features to improve the model's performance, making blurry low-light images have clear additional features. However, in contrast, the additional features mean that conventional cameras require additional equipment to acquire these features, or additional depth cameras to acquire both depth and image, which increases the cost of acquiring the image. The method mentioned in this paper is a combination of image enhancement and detection models for face recognition, which distinguishes it from generative image enhancement algorithms, and Zero-DCE's enhancement method of learning a single image without reference may be superior.

DOI: 10.54254/2755-2721/86/20241584

3. Methodology

In this section, the Zero-DCE and DSFD algorithms are proposed to assist in the task of low-light face detection, and other algorithms and models are compared to demonstrate the superior performance of these two models in this task.

3.1. Idea

Inspired by the idea of Kolmogorov-Arnold and KAN modeling, that is, any complex function or model can be composed of multiple different simple functions, we can apply this idea in the challenging task of face recognition under low-light conditions. Specifically, a complex task can be split into parts that are handled separately by multiple models, thus improving the processing efficiency and accuracy of the overall task. In this paper, we divide the face recognition task under low-light conditions into two main parts, each of which is handled by an independent model, specifically: the image enhancement part and the face recognition part.

This paper's design approach is to utilize a basic model to improve the image, and then use the improved image for face identification. The process is shown in the figure 1.

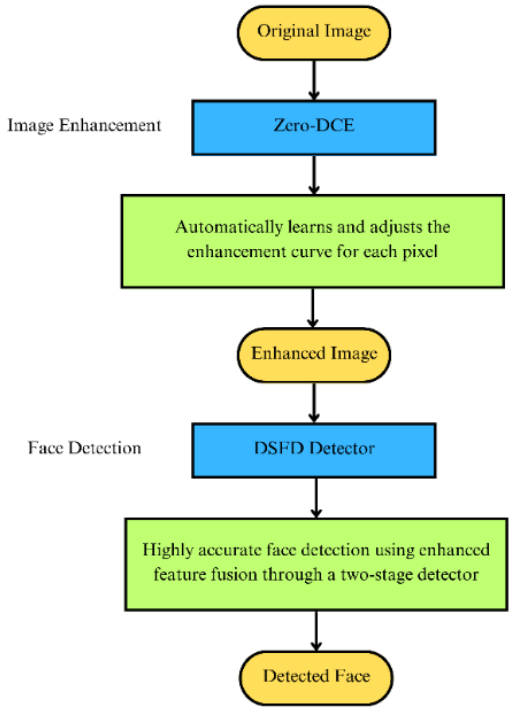

Figure 1. The flowchart of whole process. (Picture credit : Original)

3.2. Image Enhancement

Zero-DCE, or Zero-Reference Deep Curve Estimation, is a method for improving low light image that estimates depth curves without reference, whose core idea is to predict the enhancement curves directly by deep learning models, making the enhancement process fully end-to-end and automated. The primary goal of Zero-DCE is to improve low-light photos by learning image improvement curves, the method predicts the enhancement curve parameters of each pixel by a deep convolutional neural network to adaptively adapt the contrast and brightness of each image. Each pixel's enhancement curve is shown as follows:

$$I_{enhanced} = I_{input} + \sum_{i=1}^{K} \alpha_i \left(I_{input} - I_{input}^2\right) \quad (1)$$

DOI: 10.54254/2755-2721/86/20241584

Where, I_{input} is input of low-light image, $I_{enhanced}$ is enhanced image, α_i is parameter of enhancement curve, K is order of enhancement curve.

The network structure of Zero-DCE uses a lightweight convolutional neural network, whose output is the enhancement curve parameter for each pixel, and whose input is the original low-light image, which contains several convolutional layers and activation functions to determine the enhancement settings via layer-by-layer extraction of the image's characteristics.

Meanwhile, Zero-DCE uses several loss functions to ensure the quality of the image, and it lastly uses the total loss function to weight and aggregate each of the aforementioned loss functions [7].

Using the subsequent loss function, the spatial coherence loss seeks to preserve the domain surprise between the input picture and the enhanced picture to encourage the enhanced image's spatial coherence [7]:

$$L_{spa} = \frac{1}{K}\sum_{i=1}^{K}\sum_{j\in\Omega(i)}\left(\left|\left(Y_i - Y_j\right)\right| - \left|\left(I_i - I_j\right)\right|\right)^2 \tag{2}$$

where Y and I represent the local areas' average intensity values for the input picture and the improved image, and K is the number of local regions and The four domain areas centered on region i are denoted by Ω(i).

The exposure loss is meant to guarantee that the improved image's brightness values on each channel are almost at or near predetermined ideal values, which are often chosen as intermediate gray values, such 0.6 to limit overexposed areas and the exposure loss function is displayed below, indicating underexposed zones that may be controlled by adjusting the localized regions' average intensity values [7].

$$L_{exp} = \frac{1}{M}\sum_{k=1}^{M}|Y_k - E| \tag{3}$$

where Y represents the localized regions' average intensity value in the improved image, E is the optimal exposure level, which was 0.6 in the study [7], and M is the total number of 16x16 localized sections that are non-overlapping.

Based on the Gray-World assumption, the loss of color constancy is intended to restore the connection between the three adjustment channels and rectify any possible color discrepancies in the augmented image [7].

$$L_{col} = \sum_{\forall(p,q)\in\varepsilon}(J^p - J^q)^2,$$
$$\varepsilon = \{(R,G),(R,B),(G,B)\} \tag{4}$$

where J^p represents the average channel intensity value in the enhanced image for channel p, and (p, q) denotes a pair of channels.

For every curve parameter map, the following illumination smoothing loss is introduced in order to preserve a monotonic connection between surrounding pixels [7]:

$$L_{tvA} = \frac{1}{N}\sum_{n=1}^{N}\sum_{c\in\xi}\left(|\nabla_x A_n^c| + |\nabla_y A_n^c|\right)^2 \tag{5}$$

where ∇_x and ∇_y indicate the gradient operation in the horizontal and vertical directions, respectively, and N is the number of iterations [7].

With the above end-to-end learning approach, lightweight network structure and multi-loss function design, Zero-DCE maintains efficient performance while generating high-quality augmented images, providing a quality input for face detection.

DOI: 10.54254/2755-2721/86/20241584

3.3. Face Detection

DSFD consists of two main phases to improve face detection accuracy and stability through a two-stage detection framework combining FEM (Feature Enhancement Module) and PAL (Progressive Anchor Loss). Specifically, DSFD uses smaller anchors in the first stage for initial detection and larger anchors in the second stage for fine detection.

FEM enhances the feature representation by combining multi-scale features and dilated convolution (DEC), which makes the feature map more spatially discriminative and robust. Feature improvement via dilated convolution is carried out by FEM using the current layer's features as well as the higher layer's features [8], using the following formula:

$$\begin{aligned} ec_{(i,j,l)} &= f_{concat}\left(f_{dilation}\left(nc_{(i,j,l)}\right)\right) \\ nc_{(i,j,l)} &= f_{prod}\left(oc_{(i,j,l)}, f_{up}\left(oc_{(i,j,l+1)}\right)\right) \end{aligned} \tag{6}$$

Where $ec_{(i,j,l)}$ is the enhanced feature unit, $nc_{(i,j,l)}$ is the original feature unit, $f_{dilation}$ is the dilation convolution operation, f_{prod} is the elemental product operation, and f_{up} is the upsampling operation [8].

PAL optimizes the detection performance by designing progressive anchor sizes for different layers and stages. In the first stage, smaller anchors are used for initial detection; in the second stage, larger anchors are used for fine detection [8] with the following formula:

$$L_{SSL}(p_i, p_i^*, t_i, g_i, a_i) = \frac{1}{N_{conf}}\left(\sum_i L_{conf}(p_i, p_i^*) + \frac{\beta}{N_{loc}}\sum_i p_i^* L_{loc}(t_i, g_i, a_i)\right) \tag{7}$$

Where N_{conf} and N_{loc} denote the number of positive and negative anchor points, and the number of positive anchor points, respectively, L_{conf} and L_{loc} are the classification loss (softmax loss) and localization loss (smooth L1 loss), p_i, p_i^* are the predicted and true values, t_i, g_i are the actual bounding box, the anticipated bounding box, and the anchor point, a_i [8].

DSFD achieves high accuracy detection of multi-scale faces while maintaining efficient performance through techniques such as two-stage detection, feature enhancement and progressive anchor point loss. This makes DSFD highly practical in real applications. With these techniques, DSFD performs well in face detection tasks in a variety of complex scenarios and is suitable for face detection applications that require high accuracy and robustness [9,10].

4. Results and Discussion

4.1. Dataset

The Dark Face dataset from Kaggle was utilized in this investigation, which contains 6000 actual low-light photos captured at night in venues such as school buildings, streets, bridges, parks, etc., as well as the corresponding 6000 files containing labels with face location information. In this paper, it is used as a test set and using the whole test set, the efficacy of the modeling approach suggested in this research is examined.

4.2. Evaluation indicators

According to a large number of experiments, the face detection model used in this paper rarely recognizes non-face samples as faces on this dataset, and the vast majority of the results Precision is close to 1, i.e., most of the faces detected by the model are correct. In this paper, we will use recall as an important evaluation index, quantitatively analyze the performance of the above mentioned indexes on this dataset by comparing the performance of different models.

DOI: 10.54254/2755-2721/86/20241584

4.3. Image enhancement result discussion

Based on the visual comparison (Figure 2,Figure 3 and Figure 4), it can be seen that the GAN-generated image has the most vivid and bright colors, but there is more noise in the dark areas, and even color distortion, resulting in some areas of the color has deviated from the original color. The Zero-DCE-generated image is the most natural, with the least amount of noise, fewer areas of color distortion, and the details retained are more intact, but the overall brightness and color vividness may not be as bright as the GAN-generated image. However, its overall brightness and color vividness may not be as bright and vivid as the GAN-generated image. Histogram equalization generates pictures with more noise, greater loss of color, lower overall visibility of the picture, and overall performance and effect are lower than the effect of the previous two models. However, the specific prediction accuracy may not depend on the degree of visibility of the picture, but on the specific effect of the model after accepting that picture, so further effects have to be tested for the specific recall in the whole dataset.

Figure 2. Enhanced Image produced by Enlighten GAN (Picture credit : Original)

Figure 3. Enhanced Image produced by Zero-DCE (Picture credit : Original)

Figure 4. Enhanced Image produced by Histogram Equalization (Picture credit : Original)

DOI: 10.54254/2755-2721/86/20241584

4.4. Face Detection

In this paper, the three methods mentioned above are applied to the RetinaFace model and DSFD model in the Dark Face dataset, and four sets of experiments are conducted to obtain four sets of average recalls (Figure 5).

The experimental findings indicate that each image enhancing procedure in the DSFD model improves the model's recall to some extent, which is 3.29%, 5.47% and 7.1% respectively compared to the DSFD model without any processing. In the RetinaFace model, on the other hand, each enhancement algorithm has even smaller or negative enhancement, reaching only 1.03%, -0.51% and 3.03%.

According to the results, it can be seen that the images enhanced by Zero-DCE exhibit higher accuracy rates than the other methods in both DSFD and RetinaFace models, and the DSFD model outperforms the other methods in each enhancement method. The DSFD model has a higher recall than the RetinaFace model in each enhancement method, especially in the Enlighten GAN method, the DSFD model has almost 8% higher recall compared to the RetinaFace model, which indicates that the DSFD model also performs relatively well in face detection in low-light images.

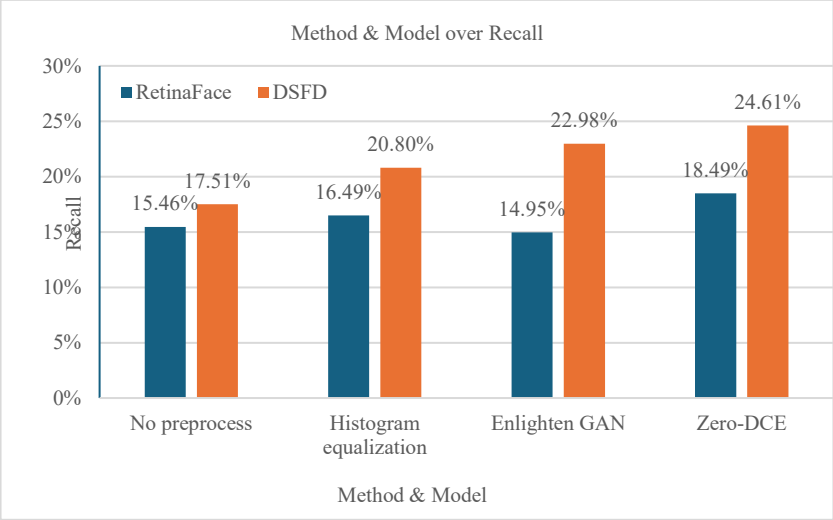

Figure 5. Method & Model over Recall (Picture credit : Original)

5. Conclusion

In this paper, the feasibility of using DSFD model for face recognition in low-light environment images by using Zero-DCE model through multiple models is verified through comparative experiments. The results show that there are varying degrees of improvement for face detection using image enhancement algorithms, and that generative image enhancement algorithms may be less effective.

These findings suggest that the staged use of multi-model co-processing tasks can be used as a more effective way to deal with complex problems, which has implications for research in the area of complex task simplification. This study does have several drawbacks, though. Firstly, the generalizability of the results may be affected by the small sample size of the dataset. The second is that the face markers of some of the data in the dataset may be too small, resulting in the model having no way to capture the facial features, which in turn cannot be recognized as faces.

This study can be extended in several ways for this complex task in extreme environments. The regression in performance indicates the division of the model into too many modules, as the problem of exchanging information between the models may cause degradation in performance and lead to long running time. So based on this, future low-light human detection may be a task of detecting the human target first, and then using a specially trained model that can perform face detection within the cropped human frame for the detection task, which may greatly increase accuracy of the existing low-light human detection task.

DOI: 10.54254/2755-2721/86/20241584

References

[1] Ren S., He K., Girshick R., Sun J., "Faster R-CNN: Towards Real-Time Obj. Det. with Reg. Prop. Nets," in IEEE Trans. Pattern Anal. Mach. Intell., vol. 39, no. 6, pp. 1137-1149, 1 June 2017.

[2] Wang J., Yang P., Liu Y., Shang D., Hui X., Song J., Chen X., "Res. on Imp. YOLOv5 for Low-Light Env. Obj. Det.," Electronics, vol. 12, iss. 14, art. 3089, 2023…

[3] Mpouziotas D., Mastrapas E., Dimokas N., Karvelis P., Glavas E., "Obj. Det. for Low Light Images," in 7th SEEDA-CECNSM, Ioannina, Greece, 2022, pp. 1-6…

[4] Kim W., "Low-Light Img. Enhanc.: A Comp. Rev. and Prospects," in IEEE Access, vol. 10, pp. 84535-84557, 2022.

[5] Lin T.-Y., Chiu C.-T., Tang C.-T., "RGB-D Based Mult.-Mod. Deep Learn. for Face Ident.," in ICASSP 2020, Barcelona, Spain, 2020, pp. 1668-1672.

[6] Thaker K., Chennupati S., Rawashdeh N., Rawashdeh S. A., "Multispectr. Deep Neural Netw. Fusion Meth. for Low-Light Obj. Det.," J. Imaging, vol. 10, iss. 1, art. 12, 2024,

[7] Guo C., Li C., Guo J., Loy C. C., Hou J., Kwong S., Cong R., "Zero-Ref. Deep Curve Estim. for Low-Light Img. Enhanc.," arXiv.org, Jan. 19, 2020.

[8] Li J., Wang Y., Wang C., Tai Y., Qian J., Yang J., Wang C., Li J., Huang F., "DSFD: Dual Shot Face Det.," arXiv.org, Oct. 24, 2018.

[9] Wei Z., Li H., Wang Z., "RetinexNet: A Low Light Image Enhancement Network Based on Retinex Theory," In 2020 IEEE/CVF Conf. Computer Vis. and Patt. Recog. (CVPRW), 2020, pp. 216-225.

[10] Jiang Z., Wang Y., Wang Y., "EnlightenGAN: Low-Light Image Enhancement via Conditional Generative Adversarial Networks," In IEEE/CVF Conf. Computer Vis. and Patt. Recog. (CVPR), 2020, pp. 7229-7238.