Hindawi
Wireless Communications and Mobile Computing
Volume 2023, Article ID 9898456, 1 page
https://doi.org/10.1155/2023/9898456

Retraction

Retracted: A Lightweight Face Verification Based on Adaptive Cascade Network and Triplet Loss Function

Wireless Communications and Mobile Computing

Received 17 October 2023; Accepted 17 October 2023; Published 18 October 2023

This article has been retracted by Hindawi following an investigation undertaken by the publisher [1]. This investigation has uncovered evidence of one or more of the following indicators of systematic manipulation of the publication process:

(1) Discrepancies in scope
(2) Discrepancies in the description of the research reported
(3) Discrepancies between the availability of data and the research described
(4) Inappropriate citations
(5) Incoherent, meaningless and/or irrelevant content included in the article
(6) Peer-review manipulation

The presence of these indicators undermines our confidence in the integrity of the article's content and we cannot, therefore, vouch for its reliability. Please note that this notice is intended solely to alert readers that the content of this article is unreliable. We have not investigated whether authors were aware of or involved in the systematic manipulation of the publication process.

Wiley and Hindawi regrets that the usual quality checks did not identify these issues before publication and have since put additional measures in place to safeguard research integrity.

We wish to credit our own Research Integrity and Research Publishing teams and anonymous and named external researchers and research integrity experts for contributing to this investigation.

The corresponding author, as the representative of all authors, has been given the opportunity to register their agreement or disagreement to this retraction. We have kept a record of any response received.

References

[1] J. Lin, C. Ye, W. Liu et al., "A Lightweight Face Verification Based on Adaptive Cascade Network and Triplet Loss Function," *Wireless Communications and Mobile Computing*, vol. 2022, Article ID 3017149, 10 pages, 2022.

Hindawi
Wireless Communications and Mobile Computing
Volume 2022, Article ID 3017149, 10 pages
https://doi.org/10.1155/2022/3017149

Research Article

A Lightweight Face Verification Based on Adaptive Cascade Network and Triplet Loss Function

Jianhong Lin,[1,2] Chaoyang Ye,[3] Weinan Liu,[4] Siqi Ren,[5] Ye Wang,[5] Wenrui Ma,[5] Bin Xu,[5] and Yifan Ding[2]

[1]*College of Computer Science and Technology, Zhejiang University, Hangzhou 310027, China*
[2]*Zhejiang Ponshine Information Technology Co., Ltd., Hangzhou 311100, China*
[3]*National (Hangzhou) New-Type Internet Exchange, Hangzhou 310009, China*
[4]*Business & Tourism Institute, Hangzhou Vocational & Technical College, Hangzhou 310018, China*
[5]*School of Computer and Information Engineering, Zhejiang Gongshang University, Hangzhou 310018, China*

Correspondence should be addressed to Siqi Ren; rensiqi@zjgsu.edu.cn

Received 8 December 2021; Revised 24 December 2021; Accepted 28 December 2021; Published 20 January 2022

Academic Editor: Liqin Shi

In the past few years, with the continuous breakthrough of technology in various fields, artificial intelligence has been considered as a revolutionary technology. One of the most important and useful applications of artificial intelligence is face detection. The outbreak of COVID-19 has promoted the development of the noncontact identity authentication system. Face detection is also one of the key techniques in this kind of authentication system. However, the current real-time face detection is computationally expensive which hinders the application of face recognition. To address this issue, we propose a face verification framework based on adaptive cascade network and triplet loss. The framework is simple in network architecture and has light-weighted parameters. The training network is made of three stages with an adaptive cascade network and utilizes a novel image pyramid based on scales with different sizes. We train the face verification model and complete the verification within 0.15 second for processing one image which shows the computation efficiency of our proposed framework. In addition, the experimental results also show the competitive accuracy of our proposed framework which is around 98.6%. Using dynamic semihard triplet strategy for training, our network achieves a classification accuracy of 99.2% on the dataset of Labeled Faces in the Wild.

1. Introduction

Artificial intelligence is one of the hottest topics in computer science which studies theories and methods used to make machine as intelligent as human beings. With the continuous breakthrough of technology in various fields, artificial intelligence has been considered as a new revolutionary technology to make progress in scientific, technological, and industrial revolution. Driven by a large amount of data, better artificial intelligence algorithms, and more powerful computing equipment, a variety of artificial intelligence applications have been used in both industries and people's daily life. Artificial intelligence industry refers to an industry that provides intelligent products and technical services to the society based on artificial intelligence technology. It has derived various new intelligent applications through enabling manufacturing, agriculture, medical, and other industries, such as subtopic detection [1], intelligent agriculture [2], and smart grid [3]. These applications have become a new engine to promote high-quality economic and social development.

With the continuous development of mobile phone applications, digital image processing and recognition have attracted more and more attention. Before the development of artificial intelligence technology, image recognition was mainly based on statistical decision-making and template matching. These traditional image recognition methods have their own limitations, and in the process of image recognition, it is necessary to manually preprocess the image and extract relevant features. If the image to be recognized has

large deformity or strong noise interference, the traditional recognition methods cannot get the expected results, resulting in poor accuracy of image recognition. With the development of artificial intelligence technology, various deep learning methods have raised image recognition technology to a new level, which has greatly improved both accuracy and real-time efficiency. The deep learning method represented by convolutional neural network, relying on its self-learning ability and computer computing ability, has achieved very good image recognition results in various application fields.

Face recognition is an important research field of digital image processing and recognition. Because of its wide application in business and security fields, face recognition becomes more and more important. For example, in the context of national fitness, people have more opportunities to enter the Asian Games venues for exercise. If efficient face recognition machines are configured in these venues, users can quickly get in and out of the sports venues. Face is the biometric information of users, which can be bound with the account to better understand the data of users' subsequent fitness, so as to further guide users to better carry out national fitness. Another useful scenario for face recognition is mobile payment. Through the face payment functionality of Alipay (the payment application of Alibaba group), users can complete payment conveniently and quickly, all of which benefit from efficient and safe face recognition algorithm. Another technique popularizes the application of face recognition is Internet of Things (IoT). IoT is one of the important application areas of 5G and future wireless communication systems. Backscatter communication can realize the low power consumption information transmission of IoT. Face recognition is an important application of IoT. The lightweight face recognition framework can be widely used in low-power devices to enrich the application scenarios of IoT. The lightweight face recognition framework combined with backscatter communication can in turn better popularize IoT. Face recognition detection and verification usually go through within two stages. One is to detect face, including face detection and face alignment, which has important practical value and significance, and has made a lot of research results [4]. The second stage is face classification. There are still many problems and challenges in this stage of research. For example, (1) face has strong variability. In different environments, the skin color of face will change due to the influence of environment. The first challenge requires face detection approach to be applicable in different scenarios; (2) the variability of face position is due to the fact that face can exist in any position in picture space or appear in a picture of any size. The second challenge requires face detection approach to check out as many faces as possible in real application. With the continuous progress of deep learning, the research heat of face recognition algorithm is rising again. Compressed convolutional neural network can complete real-time high-quality face detection on mobile platform [5]. Cascade convolutional neural network (CNN), which belongs to deep convolutional neural network (DCNN), can detect face more quickly by relying on lightweight module [6].

The main contributions of this paper can be described as following:

(1) A triplet loss function and a neural network are presented, and base on them, a lightweight face detection approach is constructed. An adaptive scale selection mechanism in first stage of the proposed face detection approach is proposed to avoid prohibitive computation which makes the approach efficient

(2) Our proposed approach achieves competitive accuracy on the dataset of Labeled Faces in the Wild (LFW) while keeps real time performance

(3) Our solution has the advantage of being lightweight and can be widely used in IoT scenarios. By incorporating the encrypted face authentication information to improve the identity authentication protocols and achieve the goal of personalized privacy protection, our approach can be applied into security field

The organization of this paper is as follows. Section 2 presents the related work. Section 3 introduces the building blocks including relevant parameters and cascade CNN. Section 4 introduces the network structure, including model training, training definition, and triple and training method selection. Section 5 shows the experimental results, including the experimental results of self-built database and LFW training set. Section 6 summarizes the paper.

2. Related Work

Before the blooming of deep learning, the performance of traditional face recognition task is advanced by handcrafted features or adjusted parameter, such as the famous local binary pattern (LBP) [7] and SIFT [8]. Ahonen et al. presented LBP texture feature-based face representation approach, in which the face features are extracted according to the LBP feature distributions and then concatenated into one single vector. Lowe paper proposed an approach for image feature extraction. The approach transforms image data into scale-invariant coordinates which can be used to extract local features. Therefore, this approach is be distinguished by Scale Invariant Feature Transform (SIFT). However, as these types of traditional methods usually take advantages of shallow network, the accuracy is relatively low. For example, the LBP can only obtain 95.17% in terms of accuracy on LFW.

With the development of CNN [9] and ImageNet, the research on face detection is on the rise again. Currently, face detection algorithms usually are based on cascade structure. For example, Mathias et al. addressed the face detection issue and presented an approach which takes advantage of deformable part model (DPM) and enjoy good performance [10]. However, as the approach requires annotations on the training data set, it suffers from large overhead of computation. Sun et al. proposed a new face detection approach which combines faster region-based CNN (RCNN) framework and a variety number of strategies, such as feature

concatenation [11]. The approach achieves remarkable performance on Face Detection Data Set (FDDB) benchmark and becomes one of the state of the art approaches in the aspect of receiver operating characteristic curves. Shi et al. pointed out that with a progressive calibration network (PCN), it is easily to distinguish face frames from nonface frames, and based on this novel PCN, they presented a rotation-invariant based face detection approach [12]. Liu et al. proposed an object detection method which discretizes the output space of bounding boxes in the image and rearranges them into a set of default boxes. The approach requires only a single deep neural network (DNN) which is faster and more accurate than the famous You Only Look Once approach [13].

In the past few years, the face detection approaches have improved. The previous methods are like the one in [14] while the new approaches take DCNN into consideration. The cascade CNN is among the most used and researched neural network in the area of face detection. To address both the effective and accurate issues in real-world face detection applications which usually have large visual variations and require discriminative detection, Li et al. proposed a cascade CNN-based face detection approach which achieves remarkable detection capability and also enjoy good performance [15]. The approach can detect the background regions at the first fast stage with low resolution and rejects these detected parts. Then, in the second stage, the approach checks high resolution part in the image to select the possible candidates for face detection. Dong and Wu focused on the Gaussian distribution and presented a face alignment approach which is based on Adaptive Cascade Deep Convolutional Neural Networks (ACDCNN) [16]. According to the Gaussian distribution among the image blocks, the approach can dynamically select the most relevant training blocks, taking advantage of an adaptively cascade CNN structure, with which, the approach enjoys high performance in accuracy, low complexity in model structure, and high robustness. To address the task handling with extensive facial landmark localization, traditional convolutional network becomes insufficient. Therefore, Zhou et al. proposed a novel approach with four-level cascade CNN [17]. Each level in the cascade CNN can predict position and rotation angles of specific image blocks and generate a coarse-to-fine detection way. Besides, this approach has the ability to process video streams immediately. To estimate the apparent age, Chen et al. proposed an approach combining a coarse-to-fine strategy and an error correction module [18]. The approach is also based on DCNN, and the used DCNN has the ability to classify the age of a detected face and can obtain a fine-grained age which further will be corrected with the error check module. The approach is relatively complex, but the performance is very good. The classic CNN-based face detection method simply stacks different types of filter layers where shallower filters can effectively check out simple non-ace samples, while deeper filters can distinguish face blocks from nonface blocks which are difficult to detect. Zhou et al. proposed a data routing mechanism that allows different layers to pass different types of samples and introduced a dual-stream context CNN architecture, which adaptively uses body part information to enhance face detection [19]. Based on them, the authors proposed an inside cascaded structure-based face detection approach where there are different classifier layers in the same CNN. The approach achieves good results in the challenging FDDB and WIDER FACE benchmark tests. Aiming at simultaneously handling four types of task, i.e., face detection, landmarks localization, pose estimation and gender recognition, Ranjan et al. presented a DCNN-based approach, i.e., HyperFace [20]. In addition, two variants based on HyperFace were proposed. The former is HyperFace-ResNet which uses the idea of residual network [21] and enjoys high performance. The latter is called Fast-HyperFace with the import of a high speed face detector. Both the two methods achieve competitive scores in the four tasks.

Usually, face detection approach needs to operate a large number of images and requires high computation devices, such as GPU cluster [22]. Guo et al. proposed an elaborately designed CNN-based face detection approach, which operates on the complete feature maps and is fast in the detection speed [23]. The authors conducted some experiments which illustrates that the approach works well on popular datasets. To better detect faces in images with nonface inputs and low-quality faces, Yu et al. proposed a novel face detection approach based on uncertainty prediction and the L2-norm of features which can reliably detect face elements from out of distribution samples and enjoys good performance [24]. The detection of small face based on DCNN usually suffers from low performance, and Ke et al. proposed a regional cascade multiscale detection approach to solve this issue [25]. The approach is made of one global face detector and some local face detectors. The product generated by the former detector on the original training set will be delivered into the latter local detectors; with this mechanism, the approach enjoys high performance. As cascade face detectors fail to achieve high accuracy and the performance of anchor-based face detectors highly depends on pretrained dataset, Yu and Tao proposed a face detection framework with efficient anchor cascade [26]. The framework takes advantage of contextual information and enjoys both efficiency and accuracy on face detection task. The experimental result shows it work better than the popular MTCNN [27].

In recent face verification algorithms, Hermans et al. compared the effects of triplet with its variant on the results [28]. Florian et al. presented a face detection system which is based on a compact Euclidean space to map face information and compute the face similarity [29]. The system which is called FaceNet utilizes triplets of face patches based on the method in [30] and achieves state-of-the-art face detection performance on the LFW dataset. Deng et al. proposed an Additive Angular Margin Loss (ArcFace) to obtain highly discriminative features, with geometric interpretation, for face recognition [31]. Lu et al. proposed the Deep Coupled ResNet (DCR) model, the backbone network was used to extract robust features that are resolution invariant, and the coupled mapping (CM) loss function was proposed to optimize the model parameters of the two branches,

respectively, on high- and low-resolution pictures [32]. Xi et al. proposed an alternating training regimen to achieve less biased classifiers and more discriminative feature representation [33]. Yu et al. used the binarization image denoising method to vanish complexity of locating feature pts, which can accurately extract facial features [34].

The application of face verification also attracts the attention of researchers. Lightweight face verification approaches be widely used in IoT scenarios, such as intelligent transportation [35] [36], especially in traffic flow prediction [37], Android applications [38], and AI-supported IoT systems [39]. To improve the identity authentication protocols and achieve the goal of personalized privacy protection, face verification approaches can be applied into security field, such as in browsers [40], social platforms [41], and cloud computing [42] by incorporating the encrypted face authentication information.

Liu et al. proposed a novel approach which takes advantage of a modified cascade CNN [43]. The proposed approach is made of three stages when training face dataset. Aiming at achieving fast face detection and higher accuracy, we introduce a triplet loss function in Section 3.1 and novel network architecture in Section 4 which constructs a new face detection approach. Using dynamic semihard triplet strategy for training, our network achieves a classification accuracy of 99.2% on the LFW dataset.

3. Building Blocks

In this section, the building blocks are presented, which consists of two parts. The former is related parameters including intersection over union, nonmaximum suppression, classifier, loss function, and triplet loss. The latter presents a three stages cascade CNN and an adaptive scale selection mechanism.

3.1. Relevant Parameters

3.1.1. Intersection over Union. Intersection over Union (IoU) is a concept used in target detection which calculates the overlap rate between the "predicted border" and "true border," i.e., the ratio of their intersection to union. Equation (1) shows the definition of IoU.

$$\mathrm{IoU} = (A \cap B)/(A \cup B). \tag{1}$$

3.1.2. Nonmaximum Suppression. The essence of nonmaximum suppression (NMS) is to search for local maximums and suppress nonmaximum elements, and IoU is used to compute NMS. When performing face detection, a window sliding method is generally adopted to generate a lot of candidate frames on the face image, and then these candidate frames are feature extracted and sent to the classifier, usually a cascade CNN. Generally, a score will be calculated on each detected face block or box. As there will be scores on many boxes, all these scores will be sorted, and one box with the highest score will be selected according to the degree of overlap, i.e., IoU, and between other frames and the current frame. In addition, the target box will not be selected if the degree of IoU is greater than a certain threshold. And except for the boxes exceeding the threshold, the high-scoring frame is selected as the detected face.

3.1.3. Classifier and Loss Function. Classifier is a function or model which conducts some mapping operations and put one item into one category. Classifier which can be applied to data prediction application is a general term to define method with classifying functionalities, such as decision trees, logistic regression, and neural network. Loss function is used to evaluate the difference between the predicted value of the classifier and the true value.

In this paper, we define two loss functions. The first one is used by the classifier, i.e., our CNN network, while the second one is used to detect face frames.

(a) The first loss function uses confidence map and bounding regression map to conduct the training job, and crossentropy is used as the loss function which is defined as Eq. (2)

$$H(y) = -\sum_{i=1}^{n} y_i' * \log(y_i), \tag{2}$$

where y_i is the predicted label which is calculated by the neural network, y_i' represents the true value of one image which is labeled in the dataset, and i is the number of elements in the dataset.

The reverse derivation of $H(y)$ is used to obtain the partial differentiation of the weights of different neural network layers.

(b) The second loss function is to address the regression issue in the task of frame detection, and we use Euclidean loss function for border regression which calculates the distance between the predicted value $y_n{}^{\wedge}$ and the label value y_n. The Euclidean loss function, $y_n{}^{\wedge}$, and y_n are defined in Eqs. (3)–(5), respectively

$$L = \frac{1}{2n}\sum_{n=1}^{N}\left\|\hat{y}_n - y_n\right\|_2^2, \tag{3}$$

$$\hat{y} = \left(x_1^{\mathrm{det}}, y_1^{\mathrm{det}}, w^{\mathrm{det}}, h^{\mathrm{det}}\right), \tag{4}$$

$$y = \left(x_1^{gt}, y_1^{gt}, w^{gt}, h^{gt}\right). \tag{5}$$

The elements in the tuple of $(x_1^{\mathrm{det}}, y_1^{\mathrm{det}}, w^{\mathrm{det}}, h^{\mathrm{det}})$ in Eq. (4) represent the x and y coordinates and height and width of the predicted face detection box while the elements in the tuple of $(x_1^{gt}, y_1^{gt}, w^{gt}, h^{gt})$ in Eq. (5) represent the x and y coordinates and height and width of the correct box in the face image.

3.1.4. Triplet Loss. In our proposed approach, the output of the cascade neural network will be the input of the triplet

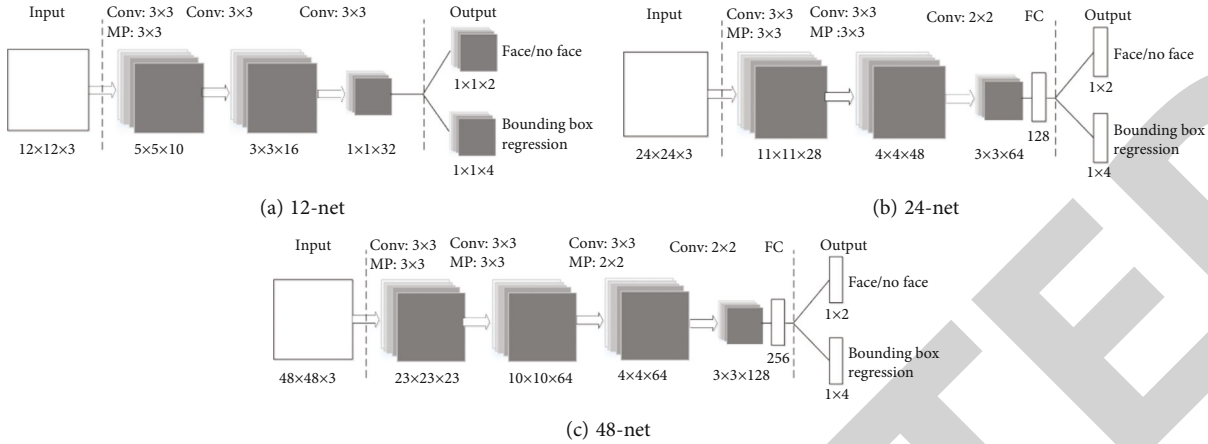

Figure 1: Neural networks.

loss function, which is an embedding mapping function represented as $f(x) \in R^d$. The triplet loss maps an image x into a d-dimensional Euclidean space. In addition, L2-normalization is used to make sure its coordinates locate on a unit hypersphere. Furthermore, to ensure that an image x_i^a of a specific person is more similar to his other image x_i^p than that of any image x_i^n of other people, the following loss function is defined as Eq. (6)

$$\text{Loss} = \sum_i^N \left[\left\| f(x_i^a) - f(x_i^p) \right\|_2^2 - \left\| f(x_i^a) - f(x_i^n) \right\|_2^2 + \alpha \right], \quad (6)$$

where $\|f(x_i^a) - f(x_i^p)\|_2^2$ is the distance between image x_i^a and image x_i^p in a d-dimensional Euclidean space, denoted by $d(a, p)$, and similarly, $\|f(x_i^a) - f(x_i^n)\|_2^2$ is the distance between image x_i^a and image x_i^n, denoted by $d(a, n)$. Additionally, the superscript a means anchor, p means positive, and n means negative.

3.2. Cascade CNN. The cascade CNN consists of two components, i.e., neural network and adaptive scale selection mechanism. Three types of neural networks are used in our proposed approach. In addition, a selection mechanism is used to decide type of neural networks to apply.

3.2.1. Neural Networks. Figure 1 shows the three types of neural network. Figure 1(a) shows structure of 12-net, Figure 1(b) shows structure of 24-net, and Figure 1(c) shows structure of 48-net. Each neural network includes one input with three parameters, i.e., width, height and channel, hidden layers, and one output. The size of network is determined by the input image size. The hidden layers are generated through two types filters, i.e., convolutional filter (Conv) and max-pooling filter (MP), each of which contains different sizes. Note that FC is full connection layer.

The first stage of the cascade CNN is a 12-net. The output is a feature map with size $1 \times 1 \times 32$ in 12-net which will further be calculated into two tensors. One is a confidence map with size $1 \times 1 \times 2$ which shows whether there exists a face or not in the input image. And the other is whether there exists a face or not in the input image. And the other is the bounding box regression with size $1 \times 1 \times 4$ which shows how the window should be adjusted in size and orientation to get a candidate frame if the input image contains a face. An adaptive scale selection mechanism is used in this stage to obtain all the candidate frames, which will be further input into 24-net to get more specified classification results and more accurate bounding boxes.

The second stage of the cascade CNN is a 24-net. The output is a feature map with size $3 \times 3 \times 64$ in 24-net which will further be calculated into two arrays. One is a confidence map with size 1×2 which shows whether there exists a face or not in the input image. And the other is whether there exists a face or not in the input image. And the other is the bounding box regression with size 1×4 which will be used to restrain the margin of a bounding box for the generated candidate frames. In this stage, if one candidate frame has a confidence score greater than 0.9 and NMS less than 0.7 with other candidate frames, then it will be kept in the candidate frame list which is defined as L_s and will be used in the finally stage.

The third stage of the cascade CNN is a 48-net. The output is a feature map with size $3 \times 3 \times 128$ in 48-net which will further be calculated into two arrays. One is a 1×2 confidence array and a 1×4D bounding info array. In this stage, if one candidate frame has a confidence score greater than 0.95 and NMS less than 0.7 with other candidate frames, then it will be regarded as the final outputs.

3.2.2. Adaptive Scale Selection Mechanism. To detect all the possible faces from a given image P with the pix size $(H \times W)$, usually, image pyramid is used which is made of different scales of the same image P. However, if too many scales are used, the computation overhead becomes insufferable. To solve this issue, in this paper, we propose an adaptive scale selection mechanism. Assume that there exists a scale set defined as $S = [S_1, S_2..S_n]$. With the scale S_i, the original image P can be transformed to another image P_i with resolution $(H_i \times W_i)$, where $H_i = H \times S_i$, $W_i = W \times S_i$. The new image P_i then becomes the input of the cascade CNN

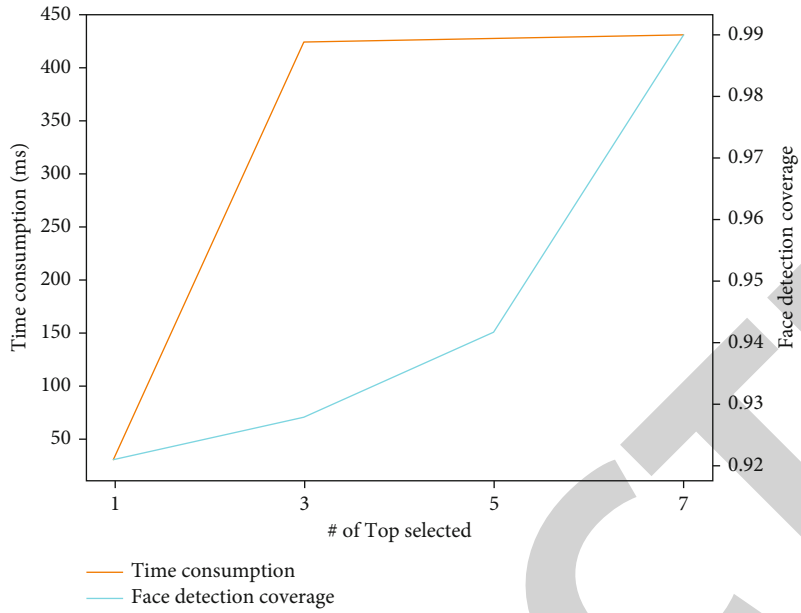

Figure 2: Relation between time consumption and coverage of found face along with the increase of top number.

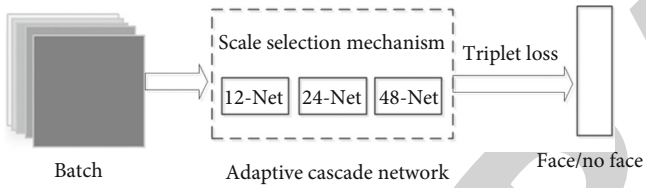

Figure 3: Network architecture.

which is described in Section 3.2. After the usage of the cascade CNN on the whole set S, we can obtain a sorted list from high value to low value. Then, we need to make a trade-off between detection speed and detection coverage. Therefore, a most appropriate number N which will be used to select top candidate frames should be decided. We conduct the experiment on the dataset and obtain the following Figure 2.

4. Network Architecture

In this section, we present the core network architecture which can be seen from Figure 3. The network includes a batch input layer and a face detection network based on adaptive cascade network. The adaptive cascade network is made of two parts. One is the three types of networks, i.e., 12-net, 24-net, and 48-net. The other is the scale selection mechanism. Both of them have been presented in Section 3.2. The last softmax layer of face detection network is replaced by a 1024-size fully connected layer (denoted as the triplet loss layer). Then, through L2 normalization, and get the embedding vectors, the triplet loss is calculated based on this feature representation.

4.1. Model Training. The introduction of triplet loss is to allow the network to learn an embedding. The network is trained using the squared L2 distances for the purpose to obtain face similarity. The face verification is completed by comparing whether the Euclidean distance of the image vector to be verified is less than a certain threshold or comparing with the known face vector in the library.

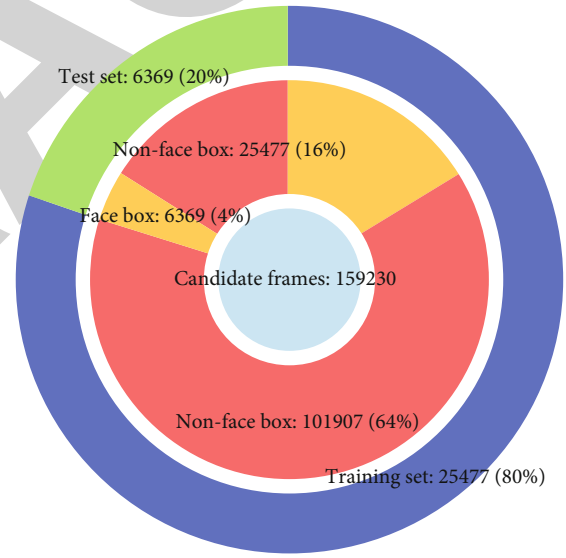

Figure 4: The construction of the self-built database.

Table 1: Approach comparison in terms of speed and accuracy.

Index	Value
CPU	Core (TM) i5-7200U 2.50GHz
Memory	8 GB
Graphics card	NVIDIA GeForce 920MX

4.2. Training Definition. Assume that a set of images input during training is in the form of$<a, p, n>$, where a and p

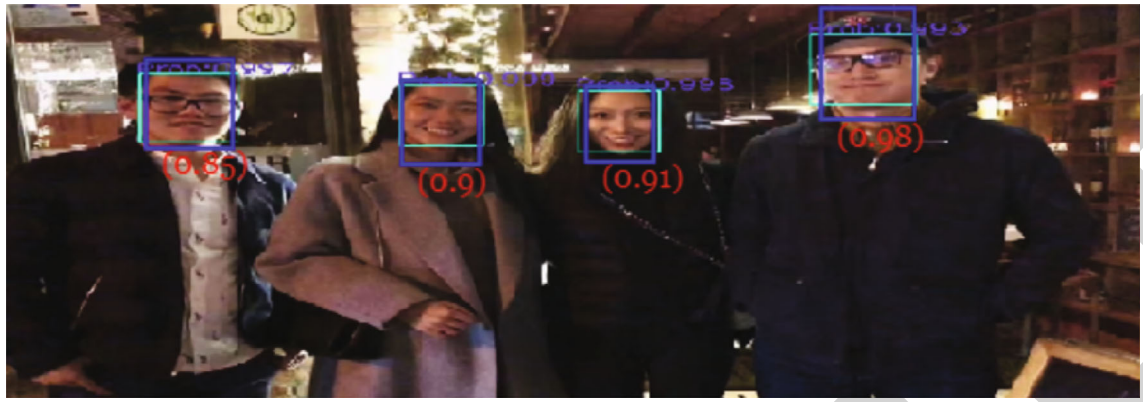

Figure 5: The detection result with and without the triplet loss function.

Table 2: Approach comparison in terms of time consumption and accuracy.

Model	Time consumption (milliseconds)	Accuracy (percentage)
Cascade network	578	93.4%
SIFT	365	95.2%
Our approach with top 1 candidate selected	31	94.5%
Our approach with top 3 candidates selected	75	96.9%
Our approach with top 5 candidates selected	152	98.6%

correspond to the same id, n corresponds to different id. The goal is to train the triplet loss layer parameters as Eq. (7).

$$d(a,p) + \alpha < d(a,n). \tag{7}$$

During the training process, the learnable parameters of all layers except triplet loss are in a frozen state, which is only used to complete feature conversion.

4.3. Triple and Training Method Selection. Since there is no target during the training, we find that there is a great influence of triplet selection of the model convergence and experimental results. Therefore, we introduce the definitions of easy triplet, hard triplet, and semihard triplet. Easy triplet: $L = 0$ is $d(a,p) + \alpha < d(a,n)$, which means that the distance between anchor and positive is less than the distance between anchor and negative. Hard triplet: $d(a,n) < d(a,p)$ means that the distance between anchor and positive is great than the distance between anchor and negative. Semihard triplet: $d(a,p) < d(a,n) < d(a,p) + \alpha$ means that the distance between anchor and positive is closely to the distance between anchor and negative.

The training method is divided into online and offline methods. The goal is to make the loss in formula (5) continue to decrease in the iterative process. The offline training method is to select all the triples in the training set and use the loss to gradient back propagation, but the distance between some anchors and negatives is very large, the calculation efficiency of using the full amount is low, and the embedding parameters cannot converge because the gradients generated by the anchors and negatives are too large; so, use online learning dynamically selects triples to solve the problems of low computational efficiency and nonconvergence of parameters.

5. Experiments

In this section, we conduct some experiments on two different face datasets. The images in the first dataset are collected from Internet while the second face dataset is the famous LFW dataset.

5.1. Performance on Self-Built Dataset. In this paper, we construct a face dataset, the images in which are all collected from Internet. We collect 12880 images totally, and based on them, 159230 candidate frames are generated. In these 159230 candidate frames, there are 31846 frames with face frames and 127384 nonface frames. The ratio of face boxes to nonface boxes is 1 : 4. In addition, in these 31846 candidate frames, the ratio of training set to test set is 8 : 2. We can see the detail construction of the self-built database from Figure 4.

The experiment is conducted in the following platform as Table 1:

We conduct the training on the self-built image dataset, the training result shows that classification accuracy of the cascade CNN achieves 99.7%, and the regression r-square is as high as 0.94. Figure 5 is the experimental result on one image of self-built image dataset with and without the triplet loss function. The green frames are without the triplet loss function, and the blue frames are with triplet loss function. We can see that the blue frames have more face contents.

Table 3: Comparison of different triplets and α.

Types	$\alpha = 0.5$	$\alpha = 1.0$
Hard triplet	97.1%	97.3%
Semihard triplet	97.5%	97.8%
Dynamic semihard triplet	99.1%	99.2%

We compare three approaches, the last of which has three variations, on the self-built dataset. We conduct this experiment 50 times and use the average as the final results.

Table 2 shows the experiments result. We can see from the result that our approach has less time consumption with all the three parameters, and the one with top 1 candidate selected works best, which only need 31 milliseconds. In addition, our approach with top 3 and 5 candidates selected work better than the common cascade network and SIFT. And the approach with top 5 candidates selected achieves a competitive score, 98.6%.

5.2. Performance on LFW. LFW dataset consists of 13233 images and 5749 individuals totally, and each image has the same resolution 250×250. Using the LFW dataset to train the embedding layer, we use different strategies of choosing triplets and different values of α to get the experimental results as Table 3. The first method is to select all hard triples to train the parameters in embedding. The second method is to select all semihard triples to train the parameters in embedding. The third method is to calculate all possible anchor-positive combinations at first, then use minibatch as the unit to calculate the distance of $d(a, p)$ in each minibatch, calculate the distance between all anchors and negative as $d(a, n)$, store them in a list, arrange the values of $d(a, n)$ in the list and select the smallest value, and calculate the distance between $d(a, p)$ and $d(a, n)$, if and only if $d(a, n) < d(a, p) + \alpha$; the a, p, n will be counted as a set of training data. When the value is 1.0, the accuracy rate for the hard triplet is 97.3%, the accuracy rate for the semihard triplet is 97.8%, and the accuracy rate for the dynamic semihard triplet is 99.2%.

6. Conclusions

This paper proposes a framework based on adaptive cascade CNN network and triplet loss for face detection and verification with fast speed and high accuracy. The framework firstly calculates the input through an image pyramid at a low resolution and adaptively selects the candidate frames. Secondly, those selected candidate frames are processed by more accurate detection network with high resolution. Finally, triple loss is calculated to conduct precise identification. The framework is very robust against complex backgrounds. We train the face verification model and complete the verification within 0.15 second for processing one image which shows the computation efficiency of our proposed framework. In addition, the experimental results also show that the competitive accuracy of our proposed framework which is around 98.6%. Using dynamic semihard triplet strategy for training, our network achieves a classification accuracy of 99.2% on the Labeled Faces in the Wild dataset.

Our future work will consider applying face verification to access control in smart grids and spatial crowdsourcing [44]. In addition, we will consider incorporating the encrypted face authentication information to improve the identity authentication protocols and achieve the goal of personalized privacy protection in face verification applications. At last, the combination of face verification and backscatter communication in IoT is another future research direction; we will consider to popularize the applications of IoT.

Data Availability

The simulation experiment data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that there are no conflicts of interest regarding the publication of this paper.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072403, 61906167, in part by the 2019 Industrial Internet Innovation Development Project-Industrial Internet Network Security Public Service Platform Project (TC190H3WN), in part by the Key Research and Development Program of Zhejiang Province under Grant 2020C01076, in part by the Natural Science Foundation of Zhejiang Province under Grant LTY21F020001 and LY21F020011, and in part by the Research Project of Zhejiang Federation of Social Sciences (2022B19).

References

[1] L. Dong, M. N. Satpute, W. Wu, and D. -Z. Du, "Two-phase multidocument summarization through content attention-based subtopic detection," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 6, pp. 1379–1392, 2021.

[2] J. Yuan, W. Liu, J. Wang, J. Shi, and L. Miao, "An efficient framework for data aggregation in smart agriculture," *Concurrency and Computation: Practice and Experience*, vol. 33, no. 10, article e6160, 2021.

[3] S. Zhao, F. Li, H. Li et al., "Smart and practical privacy-preserving data aggregation for fog-based smart grids," *IEEE Transactions on Information Forensics and Security*, vol. 16, pp. 521–536, 2021.

[4] B. Yma, A. Lw, A. Zl, and C. Fl, "A novel face presentation attack detection scheme based on multi-regional convolutional neural networks," *Pattern Recognition Letters*, vol. 131, pp. 261–267, 2020.

[5] Y. Cai, Y. Lin, L. Xia, X. Chen, and H. Yang, "Long live time: improving lifetime and security for nvm-based training-in-memory systems," *IEEE Transactions on Computer-Aided*

Design of Integrated Circuits and Systems, vol. 39, no. 12, pp. 4707–4720, 2020.

[6] M. Mathias, R. Benenson, M. Pedersoli, and L. Van Gool, "Face detection without bells and Whistles," in *European Conference on Computer Vision*, pp. 720–735, Zurich, Switzerland, 2014.

[7] D. Shi and H. Tang, "Face recognition algorithm based on self-adaptive blocking local binary pattern," *Multimedia Tools and Applications*, vol. 80, no. 16, pp. 23899–23921, 2021.

[8] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, 2004.

[9] S. Cebollada, L. Payá, X. Jiang, and O. Reinoso, "Development and use of a convolutional neural network for hierarchical appearance-based localization," *Artificial Intelligence Review*, pp. 1–28, 2021.

[10] W. Liu, D. Anguelov, D. Erhan et al., "SSD: Single shot multibox detector," in *European conference on computer vision*, pp. 21–37, Amsterdam, The Netherlands, 2016.

[11] X. Sun, P. Wu, and S. Hoi, "Face detection using deep learning: an improved faster rcnn approach," *Neurocomputing*, vol. 299, pp. 42–50, 2018.

[12] X. Shi, S. Shan, M. Kan, S. Wu, and X. Chen, "Real-Time rotation-invariant face detection with progressive calibration networks," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Salt Lake City,USA, 2018.

[13] J. Redmon, S. Divvala, R. Girshick, and A. Farhadi, "You only look once: unified, real-time object detection," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Las Vegas, NV, USA, 2016.

[14] J. Li, T. Wang, and Y. Zhang, "Face detection using surf cascade," in *IEEE International Conference on Computer Vision Workshops*, pp. 2183–2190, Barcelona, 2011.

[15] H. Li, Z. Lin, X. Shen, J. Brandt, and G. Hua, "A convolutional neural network cascade for face detection," in *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, Massachusetts, 2015.

[16] Y. Dong and Y. Wu, "Adaptive Cascade deep convolutional neural networks for face alignment," *Computer Standards & Interfaces*, vol. 42, pp. 105–112, 2015.

[17] E. Zhou, H. Fan, Z. Cao, Y. Jiang, and Q. Yin, "Extensive facial landmark localization with coarse-to-fine convolutional network cascade," in *Proceedings of the 2013 IEEE International Conference on Computer Vision Workshops*, Sydney, Australia, 2013.

[18] J. C. Chen, A. Kumar, R. Ranjan, V. M. Patel, A. Alavi, and R. Chellappa, "A cascaded convolutional neural network for age estimation of unconstrained faces," in *IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*, Niagara Falls, NY, USA, 2016.

[19] K. Zhang, Z. Zhang, H. Wang, Z. Li, Y. Qiao, and W. Liu, "Detecting faces using inside cascaded contextual CNN," in *IEEE International Conference on Computer Vision (ICCV)*, Venice, Italy, 2017.

[20] R. Ranjan, V. M. Patel, and R. Chellappa, "HyperFace: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition," *IEEE Transactions on Pattern Analysis & Machine Intelligence*, vol. 41, no. 1, pp. 121–135, 2019.

[21] K. He, X. Zhang, S. Ren, and J. Sun, "Identity mappings in deep residual networks," in *European Conference on Computer Vision*, Amsterdam, The Netherlands, 2016.

[22] J. Xu, J. Wang, Q. Qi, H. Sun, and D. Yang, "Effective scheduler for distributed dnn training based on mapreduce and gpu cluster," *Journal of Grid Computing*, vol. 19, no. 1, 2021.

[23] G. Guo, H. Wang, Y. Yan, J. Zheng, and B. Li, "A fast face detection method via convolutional neural network," *Neurocomputing*, vol. 395, pp. 128–137, 2020.

[24] C. Yu, X. Zhu, Z. Lei, and S. Z. Li, "Out-of-distribution detection for reliable face recognition," *IEEE Signal Processing Letters*, vol. 27, pp. 710–714, 2020.

[25] X. Ke, J. Li, and W. Guo, "Dense small face detection based on regional cascade multi-scale method," *IET Image Processing*, vol. 13, no. 14, pp. 2796–2804, 2019.

[26] B. Yu and D. Tao, "Anchor cascade for efficient face detection," *IEEE Transactions on Image Processing*, vol. 28, no. 5, pp. 2490–2501, 2019.

[27] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint face detection and alignment using multitask cascaded convolutional networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016.

[28] A. Hermans, L. Beyer, and B. Leibe, "In defense of the triplet loss for person re-identification," 2017, https://arxiv.org/abs/1703.07737.

[29] F. Schroff, D. Kalenichenko, and J. Philbin, "FaceNet: a unified embedding for face recognition and clustering," in *IEEE Conference on Computer Vision and Pattern Recognition*, Boston, Massachusetts, 2015.

[30] Y. Bengio, J. Louradour, R. Collobert, and J. Weston, "Curriculum learning," in *Proceedings of ICML*, New York, NY, USA, 2009.

[31] J. Deng, J. Guo, and S. Zafeiriou, "ArcFace: additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4690–4699, Long Beach, CA, 2019.

[32] Z. Lu, X. Jiang, and C. C. Kot, "Deep coupled ResNet for low-resolution face recognition," *IEEE Signal Processing Letters*, vol. 25, no. 4, pp. 526–530, 2018.

[33] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Feature transfer learning for face recognition with under-represented data," in *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, Long Beach, CA, 2019.

[34] X. Yin, X. Yu, K. Sohn, X. Liu, and M. Chandraker, "Improvement of face recognition algorithm based on neural network," in *2018 10th International Conference on Measuring Technology and Mechatronics Automation (ICMTMA)*, Changsha, China, 2018.

[35] G. Xu, X. Li, L. Jiao et al., "BAGKD: A batch authentication and group key distribution protocol for VANETs," *IEEE Communications Magazine*, vol. 58, no. 7, pp. 35–41, 2020.

[36] G. Xu, W. Zhou, A. K. Sangaiah et al., "A security-enhanced certificateless aggregate signature authentication protocol for InVANETs," *IEEE Network*, vol. 34, no. 2, pp. 22–29, 2020.

[37] W. Shu, K. Cai, and N. N. Xiong, "A short-term traffic flow prediction model based on an improved gate recurrent unit neural network," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2021.

[38] G. Xu, W. Wang, L. Jiao et al., "SoProtector: safeguard privacy for native SO files in evolving mobile IoT applications," *IEEE Internet of Things Journal*, vol. 7, no. 4, 2020.

[39] G. Xu, Y. Zhao, Y. Jiao et al., "TT-SVD: an efficient sparse decision making model with two-way trust recommendation in the

AI enabled IoT systems," *IEEE Internet of Things Journal*, vol. 8, no. 12, pp. 9559–9567, 2021.

[40] G. Xu, X. Xie, S. Huang et al., "JSCSP: a novel policy-based XSS defense mechanism for browsers," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2020.

[41] G. Xu, B. Liu, L. Jiao et al., "Trust2Privacy: a novel fuzzy trust-to-privacy mechanism for mobile social networks," *IEEE Wireless Communications*, vol. 27, no. 3, pp. 72–78, 2020.

[42] W. Shu, K. Cai, and N. Xiong, "Research on strong agile response task scheduling optimization enhancement with optimal resource usage in green cloud computing," *Future Generation Computer Systems*, vol. 124, pp. 12–20, 2021.

[43] G. Liu, J. Lin, Y. Ding, S. Yang, and Y. Xu, "A new face detection framework based on adaptive cascaded network," in *International Conference on Frontiers in Cyber Security*, Tianjin, China, 2020.

[44] S. Han, J. Lin, S. Zhao et al., "Location privacy-preserving distance computation for spatial crowdsourcing," *IEEE Internet of Things Journal*, vol. 7, no. 8, pp. 7550–7563, 2020.