The Visual Computer (2021) 37:217–226
https://doi.org/10.1007/s00371-020-01794-9

ORIGINAL ARTICLE

Face recognition in unconstrained environment with CNN

Hana Ben Fredj[1] · **Safa Bouguezzi**[1] · **Chokri Souani**[2]

Published online: 18 January 2020

Abstract
In recent years, convolutional neural networks have proven to be a highly efficient approach for face recognition. In this paper, we develop such a framework to learn a robust face verification in an unconstrained environment using aggressive data augmentation. Our objective is to learn a deep face representation from large-scale data with massive noisy and occluded face. Besides, we add an adaptive fusion of softmax loss and center loss as supervision signals, which are helpful to improve the performance and to conduct the final classification. The experiment results show that the suggested system achieves comparable performances with other state-of-the-art methods on the Labeled Faces in the Wild and YouTube face verification tasks.

Keywords Face recognition · Deep learning · Data augmentation

1 Introduction

Face recognition is a well-studied problem in computer vision. Automatic face recognition is an important vision task in several practical applications such as identity verification, intelligent visual surveillance and immigration automated clearance systems. According to different application scenarios, it can be classified into two different tasks: face verification and face identification. The former aims to determine whether a given pair of face images is from the same person or not, while the second is to recognize the person from a set of gallery face images and find the most similar one. Nevertheless, face recognition in real applications is still a challenging task [1]. The main reason is that the face is a non-rigid object, and it often has varied appearance owing to numerous facial expressions, different ages, multiple angles and more importantly the various illumination intensities. In addition, there are still many factors which affect the face recognition performance, such as occlusions and poses.

In recent years, deep learning has become more and more prevalent in computer vision. In the last decade, convolutional neural networks (CNNs) have become popular techniques for solving computer vision problems. Numerous vision tasks, such as image classification [2], object detection [3] and face recognition [4–8], have benefited from the robust and discriminative representation learned via CNNs. Indeed, a neural network can learn effective features from repeated convolution and pooling operations based on a large-scale dataset. Recently, CNNs have shown a powerful capability and have become the most effective means for dense prediction problems, especially in the field of face recognition. Some of improvement in face recognition has been propounded in the past decade, which was based on CNNs [9, 10]. The recent face recognition methods have made considerable progress, even beating human beings on the Labeled Faces in the Wild (LFW) and Wild and YouTube face (YTF) benchmarks.

To achieve optimal accuracy, the scale of the training dataset for CNNs has been consistently increasing. Data augmentation (DA) is a potential solution that artificially inflates a database by using a domain-specific synthesization to add more invariant examples [11]. In addition, it is a set of computationally inexpensive methods previously used to reduce overfitting in training a CNN [12]. However, large-scale datasets often contain massive noisy signals especially

✉ Hana Ben Fredj
ben.fredj.hanaa@gmail.com

Safa Bouguezzi
safabouguezzi@yahoo.fr

Chokri Souani
chokri.souani@gmail.com

[1] Laboratoire de microélectronique et instrumentations, Faculté des sciences de Monastir, Université de Monastir, Monastir, Tunisia

[2] Institut supérieur des sciences appliquées et de technologie de Sousse, Université de Sousse, Sousse, Tunisia

when they are automatically collected via image search engines or from movies.

Face recognition work based on CNNs has achieved ideal recognition rates, except that the majority of studies have used original data on different databases and have lacked more complex situations. Authors in some works pursued DA methods of increasing training data and to generate more examples according to the requirement of deep learning.

Our approach consists in forcing learning in different more difficult situations to solve problems in face recognition caused by data corruption, variation in illumination, occlusion and missing parts. Inspired by DA methods and to further make more complex situations in the training dataset, we use, in this paper, an aggressive DA, so as to generate more face images and to learn a deep face representation from the large-scale data within a CNN model. We propose a collection of perturbations on aligned faces. In fact, we train our model with various transformations, such as noisy regions, blurring, contrast, and variation in illuminations. Especially, we fuse the entire face with its components and occluded face. Besides, to obtain the deep features on large-scale datasets, we train a robust CNN with the joint supervision of softmax loss and center loss. The two key learning objectives, inter-class dispersion and intra-class compactness, are very essential to face recognition. According to the results of experimental analysis, our model has a good performance compared with the state of the art.

The rest of this paper is organized as follows. Section 2 reviews some related previous work. Section 3 presents the architecture network and the DA approach. The experimental setup and results are presented in Sect. 4. Section 5 offers our conclusion.

2 Related work

In this section, we make an overview of existing work on face recognition.

There are some work in the literature in the field of computer vision [13–17]. Especially, face recognition has been a prevalent research field in pattern recognition, which it consists of two stages: face detection and face recognition.

Several recognition techniques have been developed to capture discriminative features for better performances. The traditional approaches usually include two steps: high-dimensional feature extraction and a classifier design. The CNN models naturally integrate the feature extractor and the classifier in an end-to-end fashion. The face representations obtained by the methods are effective.

Newly, CNN-based applications, reminiscent of FaceNet [18], DeepFace [6] and DeepID [19], are extensively utilized in face recognition tasks and have shown necessary results in free environments. Compared with the conventional face recognition methods, face recognition models based on deep networks can always achieve better performances. For instance, FaceNet [18] operates very deep networks to perform face recognition. It utilizes approximately 8,000,000 images of 2,000,000 people. Applied to the largely used LFW database, this system achieves 99.63% of accuracy. Since 2012, Deep CNNs (DCNN) has become prevalent. This is due to the large amounts of training data and adaptable computing resources such as GPUs. For example, Krizhevsky et al. [20] trained a convolutional network to classify images in ILSVRC-2012 competition and obtained attractive recognition accuracy. Meanwhile, the DCNN architecture, such as GoogLeNet [9] and VGG [21], has been much wider and deeper, leading to enormous network parameters and good performances.

A lot of approaches have been also suggested to improve the face verification performance in an unconstrained environment, and some of them have exhibited impressive results. Guo et al. [22] propounded a deep network model which took both visible light and near-infrared images into account to perform face recognition. The experimental results demonstrated that the model was very effective in real-world scenarios and performed much better in terms of illumination change than other state-of-the-art models. The authors in [23] developed a facial expression recognition algorithm based on the deep learning method. This adaptive model parameter initialization, based on the multilayer maxout network linear activation function, allowed initializing the CNN and the long–short-term memory (LSTM) network method. The experiments showed that the facial expression recognition method would accurately identify various expressions and have a good adaptive ability. Jiang-Jing et al. [24] put forward a simple and efficient DA approach, which uses artificial landmark perturbation to generate a huge number of misaligned face images, to train DCNN model robust against landmark misalignment. The experimental LFW and YTF verify the effectiveness of the approach. The authors in [25] present a light CNN framework to learn a compact embedding on large-scale face data with massive noisy labels. The experimental results showed that the proposed framework was efficient in computational costs and storage spaces. In [26], the authors introduced a new layer to embed the patch strategy in convolutional architecture to improve the effectiveness of face representation. This approach made a better use of the interactions between global and local features in the model. Two baseline CNNs (i.e., AlexNet and ResNet) were used to analyze the effectiveness of their method. The experiments indicated that the suggested system achieves comparable performance with other state-of-the-art methods on the LFW and YTF tasks. Wen et al. [27] supervised a CNN by a novel signal center loss together with the softmax loss and obtained the

state-of-the-art accuracy on three important face recognition benchmarks.

Following the trend, we learn face features by using CNN with the joint supervision of softmax loss and center loss to improve the performance of face representation in the paper. Inspired by DA methods, we also employ an aggressive DA method to develop a CNN framework and to learn a robust face verification in an unconstrained environment.

3 Data augmentation

In real-world applications, natural data can still exist in a variety of conditions such as varying illuminations and noisy information. These latter are among the most important factors that significantly affect the performance of face recognition algorithms, and they also draw much attention in deep learning. This paper studies a CNN framework to learn a deep face representation with more difficult conditions for the appearance of faces. Hence, the DCNNs have a powerful feature extraction ability and can obtain competitive extraction by using massive training sets [28]. Therefore, if we want to solve the complex change problem utilizing deep networks, there must be enormous training data that have various scenarios. The idea is to use the DA to present the samples dataset in the different variations, as we pointed out earlier, to force the CNN framework to learn a robust face representation with different modifications. This can lead to better performances. We account for these situations by training our neural network with additional synthetic modified data. DA is a very common and important preprocessing step for CNN-based methods [24, 29] to achieve considerable performance. There are different DA methods, which are used to find various situations, such as flipping, color casting, blurring, noise, histogram and sigmoid. Moreover, illumination is an important influence factor for the dataset robustness. Several types of illumination conditions may affect the results of object detection and recognition. Also, various methods have been proposed to transform a real face image to a new type, such as pose transfer, hairstyle transfer, expression transfer, makeup transfer, and age transfer [30]. In order to improve the dataset robustness against different illumination types, variation illumination is used for DA, i.e., changing in the tone, luminance or contrast in images. Herein, a general framework for the association of face components or partial faces with a full face is introduced. The DA strategy in this work also contains a combination of faces occluded to learn data in more difficult situations in the training dataset. This method is adopted to extract features of various regions with CNN model. We aim to generate more face images with misalignment for DCNN training to improve, significantly, the face recognition rate.

4 Network architecture

DCNNs have achieved outstanding performances on image classification. There are many tricks for DCNN training with very deep architecture. This section describes the CNN used in our experiments. In this paper, we use a CNN model based on GoogLeNet style Inception models, called inception-v1 [9], which contains the most mainstream components of CNN architecture. The key idea is to devise this architecture and deploy multiple convolutions with multiple filters and pooling layers, simultaneously, in parallel within the same inception layer. Furthermore, each inception layer has filters of different sizes (e.g., 1×1, 3×3, 5×5) to find the optimal local construction. We find more details of the configuration of the model in Table 1. Each convolutional neuron treats data only for its receptive field. It is followed by additional nonlinear operations, ReLU, max (0.x) which is an activation function. Then this model uses pooling layers, which are fixed directly after the convolutional layers, whereas softmax is used as a loss function that computes the probability of the Kth output assigned to the Kth class. Softmax (or multinomial logistic) regression is a generalization of the logistic regression, and it is a kind of linear regression [31]. It is used in several problems including text classification. In our case, it is utilized for the classification of faces within the target classes through the logistic operation. More than the softmax function, we use the center loss function to train the deep model.

To develop an effective loss function and to improve the discriminative power of deeply learned features, we use the center loss function based on the idea of authors in [32]. The center loss minimizes the intra-class variations while keeping the features of multiple classes separable by softmax. Equation 1 gives the center loss function as follows:

$$l_c = \sum_{i=1}^{m} \left\| x_i - c_{yi} \right\|_2^2 \tag{1}$$

where l_c is the center loss, m is the number of training samples in a min-batch, $x_i \in R_d$ denotes the ith training sample, yi is the label of x_i. $c_{yi} \in R_d$ denotes the y_ith class center of deep features, and d is the feature dimension. The idea is to adopt the joint supervision of softmax loss and center loss to train the network, when training DCNN. Equation 2 describes this fusion as follows:

$$L = L_s + \lambda L_c \tag{2}$$

where L is the DCNN total loss, L_s is the softmax loss, L_c is the center loss, and λ is the scalar used for balancing the two loss functions. However, training the CNN with the center loss is easier than training with the triplet loss method, which is used on the FaceNet model.

Table 1 CNN model used in this paper based on GoogLeNet (inception architecture)

Type	Filter size/stride	Output size	Depth	Parameters (K)
Convolution	7×7/2	112×112×64	1	2.7
Max pool	3×3/2	56×56×64	0	
Convolution	3×3/1	56×56×192	2	112
Max pool	3×3/2	28×28×192	0	
Inception (3a)		28×28×256	2	159
Inception (3b)		28×28×480	2	480
Max pool	3×3/2	14×14×480	0	
Inception (4a)		14×14×512	2	364
Inception (4b)		14×14×512	2	437
Inception (4c)		14×14×512	2	463
Inception (4d)		14×14×528	2	580
Inception (4e)		14×14×832	2	840
Max pool	3×3/2	7×7×832	0	
Inception (5a)		7×7×832	2	1072
Inception (5b)		1×1×1024	2	1388
Avg pool	7×7/2	1×1×1024	0	
Fully connection		1×1×1000	1	10.575
Softmax		1×1×1000	0	

5 Experiments

In what follows, we first describe the details of datasets and then present the methodology of training and testing. A set of experiments are conducted on the LFW and YTF for the verification task. We also discuss the effectiveness of the joint supervision of softmax loss and center loss to train the network. We describe the influence of combination of different data augmentation methods on LFW and YTF datasets. Besides, we provide a comparison with the state-of-art methods on LFW and YTF.

In order to demonstrate the effectiveness of our proposed method of face verification, we used typical databases in this domain.

LFW It is the standard benchmark and a public dataset for automatic face verification. The database comprises 13,233 images of 5749 people, where 1680 subjects contain more than two images and 4096 subjects consist of only one image. The face images taken from LFW are taken under an unconstrained environment with face variations, such as poses, illuminations, expressions, and occlusions.

YTF This is a popular face dataset that allows evaluating face recognition performance. It contains 3425 videos of 1595 different people. Various videos were downloaded from YouTube. On average, 2 videos are available for each subject. To determine the face verification performance, 5000 pairs of face videos are used for demonstrating the face verification performance. The YTF face frames not only contain large facial variations (such as occlusions, expressions, poses and lighting), but also suffer from various levels. It is treated as a more challenging testing dataset for face verification.

CASIA-WebFace The CASIA-WebFace dataset has face images of celebrities taken from websites. It is a classic public dataset with wide face subjects, namely 10,575 subjects with 494,414 face images, and it is utilized for scientific research of unconstrained face recognition. Furthermore, it can be treated as a standard training dataset for developing face recognition methods, and it has no overlap with the YTF and LFW datasets.

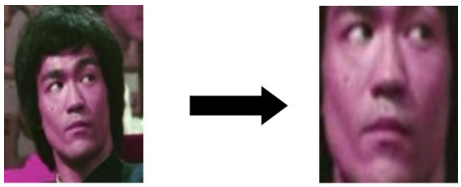
Next, we describe the details of image preprocessing for the mentioned datasets and the augmentation data methods.

5.1 Face detection and data augmentation methods

The training and testing datasets contain multiple images with numerous conditions and positions. In fact, we need an efficient method for the detection and alignment of faces. We use the multi-task cascaded CNN (MTCNN) [33]-based framework for joint face detection and alignment. RetinaFace is a powerful face detector method for face recognition applications. It could easily handle faces with pose variations, but it still had difficulty under complex scenarios [34].

The MTCNN is capable of detecting faces by locating facial landmarks (i.e., two eyes, a nose, and mouth endpoints). Figure 1 presents an example of the face image alignment from the CASIA-WebFace dataset using the MTCNN method.

As previously mentioned, our objective is to have a deep face representation of large-scale data with a massive



(a) (b)

Fig. 1 Example of face image alignment from CASIA-WebFace dataset. **a** Original image, **b** result of aligned face with MTCNN method

complex and uncontrolled environment. For that, we change the face representation by including multiple situations of information disruption in the CASIA-WebFace database before the training step. Figure 2 illustrates the different DA methods, which are gathered and used in our work.

In order to make the learning environment more difficult, we propose a collection of DA methods by perturbing original information (aligned faces). Thus, we train our model with various transformations such as flipping, histograms, noise, and blurring (Fig. 2a). In addition, we suggest training data that have various illumination intensities (Fig. 2b) to solve the illumination change problem. Afterward, we add randomly occluded faces (Fig. 2c) to learn data in more difficult situations. Besides, a mix between full faces and these components is present in this approach (Fig. 2d).

5.2 Training methodology

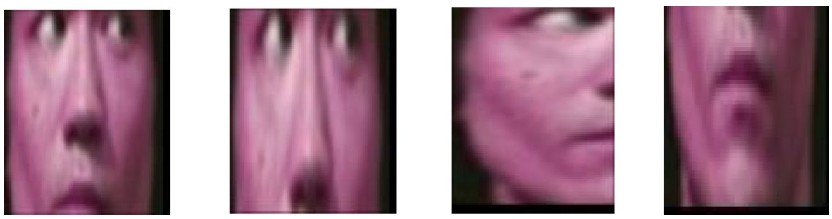
The details of the training strategy are presented as follows. The model is trained on GPU GeForce GTX 1080 Ti. We adopt a public available deep learning framework, Tensorflow, to train our model. This model is characterized by

Fig. 2 Examples generated by of different DA methods used in our work; **a** Flipping + Histogram + Noise + Blurring, **b** examples with Different illumination, **c** examples with different occluded face, **d** examples of cropped face parts

(a) Examples with the different DA methods: Flipping+Histogram+Noise+Blurring

(b) Examples with Different illumination

(c) Examples with different occluded face



(d) Examples of cropped face parts

aggressive DA and by combining the softmax loss and the center loss put forward to achieve better performances. As indicated, for face alignment, the MTCNN is used, proving that it gives very good performances for the alignment of train/test sets. We train the model on an aligned dataset for 100 epochs with an RMSProp optimizer. To improve the generalization of the CNN model, several techniques are utilized in this work. In order to perform a regulation in the loss function and to avoid overfitting, the weight decay parameter is used. The coefficient of the latter is set to $5e^{-4}$ for the convolutional layers and fully connected layers. Added to that, the model is regularized using a dropout layer. The learning rate is initially set to 0.05 and then decreases by a factor of 0.1 when the validation set accuracy stops rising. The momentum coefficient is also set to 0.9. For details of testing, the similarity score is calculated by the cosine distance of a pair of features after transforming the representation. We report the results on LFW and YTF following the standard protocol of restricted, labeled outside data.

Besides, as mentioned before in this paper, we join the softmax loss and center loss. To further demonstrate the effectiveness of the center loss, we conduct the experiments on the LFW dataset. We train on the augmented dataset CASIA-WebFace using the information disruption and test on the 6000 face pairs on LFW, using softmax only on the one hand. On the other hand, we also retrain the network under the supervision of the softmax loss and center loss. The result is provided in Table 2. We can observe that when the model is trained with the fusion of both losses, the accuracy on LFW boosts by 1.07%, compared, respectively, with the model trained with the softmax loss only.

The combination of the center loss and the softmax loss gives better results than simply using anyone of them separably.

5.3 Model trained with different DA methods

To further verify the robustness of our method, we compare the performance with different DA methods. Therefore, we mix the different DA methods and train our model as follows.

A: No data augmentation
B: Different illumination + flipping
C: Occlusion + blurring + noise + histogram + **B**
D: we have collected the different DA methods in our work: C + cropped face parts

The results of each data augmentation method on LFW and YTF datasets are provided in Table 3, and the ROC curves are shown in Figs. 3 and 4.

As shown in Table 3, for case A, without DA, the classification rate reaches only 94.5% on LFW. For B, C and D, the rate increases gradually. We can see clearly, by changing the lighting with flipping methods, that the face recognition accuracy can reach 97.7%. By adding the occlusion, blurring, noise and histogram methods, respectively, the model can achieve 98.4% rate. For case D, we make a balance by adding the cropped parts of faces. Thus, among

Table 2 Effect of center loss method by determining classification accuracy (%) on the LFW dataset

Method	Accuracy (%)
Softmax loss	98.133 ± 0.013
Softmax loss + center loss	99.2 ± 0.04

The results are recorded in two cases. First, the network is trained with the softmax loss only. Second, the network is trained with softmax loss + center loss

Table 3 Classification accuracy (%) on LFW and YTF datasets with different methods DA

Method	Accuracy (%)	
	LFW	YTF
A	94.5 ± 0.008	91.9 ± 0.011
B	97.7 ± 0.006	93.2 ± 0.014
C	98.4 ± 0.007	94.38
D	99.2 ± 0.04	96.6

A: No data augmentation, **B**: different illumination + flipping, **C**: Occlusion + blurring + noise + histogram + **B**, **D**: we have collected the different DA methods in our work: **C** + cropped face parts

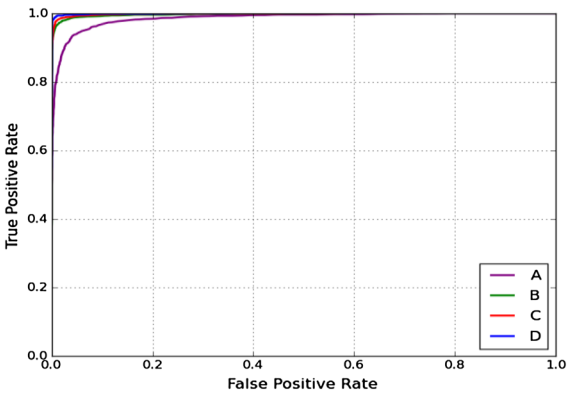


Fig. 3 Influence of combination different data augmentation methods on LFW dataset. **A**: No data augmentation, **B**: Different illumination + flipping, **C**: Occlusion + blurring + noise + histogram + **B** **D**: We have collected the different DA methods on our work: **C** + cropped face parts

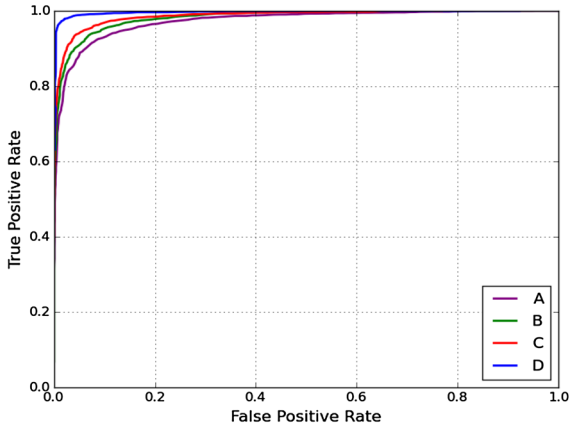


Fig. 4 Influence of combination of different data augmentation methods on YTF dataset. **A**: No data augmentation, **B**: Different illumination + flipping, **C**: Occlusion + blurring + noise + histogram + **B**, **D**: We have collected the different DA methods on our work: **C** + cropped face parts

these DA methods, case D performs much better than the other methods and it achieves the highest rates on the verification task with 99.2%. In other words, when incorporating more DA methods, the performance can be significantly improved. Actually, the fused methods achieve high recognition rate. Furthermore, adding the parts cropped of faces and the different DA combinations can still improve the performances. This test shows that the combination of the different DA methods is practical and efficient. It can be used to improve the face recognition performance in an uncontrolled and complex environment.

For the YTF dataset, it is remarkable that for a method without DA, the accuracy decreases to 91%, whereas the accuracy increases progressively with B, C and D. The significant result is obtained with method D. Thus, according to Table 3, we can also demonstrate the significance of the mix of diverse DA methods on the YTF dataset. Accordingly, on both datasets, our model is reasonable for better performances.

Figures 3 and 4 provide ROC curves for each tested augmentation technique on the LFW and YTF datasets. As shown, the training model without DA fails to attain better results (purple curves). We can easily find that the performance of the proposed method is at the top level (blue curves). The suggested method D is better than the others; the main reason is the addition of the cropped parts of faces. Besides, the different DA combination chosen in our work can still improve the performance. This method enables also the learned feature to be more discriminative. Thereafter, these results illuminate the effectiveness of the DA methods on the LFW and YTF datasets.

Table 4 Performance comparison with state-of-the-art methods on LFW

Method	Accuracy (%)
[35]	98.00
[36]	73.897
[37]	99.33 ± 0.3
Our method	99.2 ± 0.04

We achieve comparable results with other studies [35–37] based on different DA methods

5.3.1 Comparison with the state of the art

This section presents a comparison of our method with other state-of-the-art methods based on DA on LFW dataset, accompanied with a description of the multiple DA systems. Our aim consists in forcing learning in different more difficult situations. Inspired by DA methods and to further make more complex situations in the training dataset, we use, in this paper, an aggressive DA, so as to generate more face images and to learn a deep face representation from the large-scale data within a CNN model.

There are several published DA methods, which have reached important results in terms of verification rate. Table 4 provides a comparison with other studies based on different DA methods on LFW. As illustrated in this table, obviously, our method outperforms [35, 36]. Indeed, the authors in [35] described methods of enriching an existing dataset with important facial appearance variations by manipulating the faces it contained, which generated images with facial appearance variations, including poses, shapes and expressions. As given in Table 4, the effect of training and testing with synthesized images on the LFW reaches 98.00%, which is less than our work. However, this model did not take into account inter-class and intra-class distances jointly, hence neglecting the spatial structures underlying inter-class and intra-class data samples. On the other hand, the authors in [36] described a method to generate reasonable virtual samples, so as to prevent imbalance classification results. This method was based on joint Bayesian face analysis, and the experiments were conducted based on high-dimensional LBP features as well as features extracted by a shallow CNN. This method is already acceptable considering the model size and running speed, since it is trained with limited samples and reduced network parameters. The performances of this method have been limited as well. The accuracy has risen only to 71.235% and 73.897% with the augmented dataset. As a consequence, it is not much fascinating compared to known highly accurate algorithms and our work. Also, the results in [36, 37] remain in an environment, which is not rich by hard and difficult situations.

Evidently, [37] has the most significant performance. The experimental result on popular LFW for the verification rate

achieves 99.33 ± 0.39%. This is because the method presents five data augmentation methods dedicated to these factors: landmark perturbation and four synthesis methods (hairstyles, glasses, poses, illuminations), which improve the face recognition performance, hence increasing the effective size of the training set. The DA methods are easy to implement and integrate. However, the authors in [37] have not make the environment more difficult of the training set since they have used the most common occlusion method, as wearing glasses. In addition, the generalization ability of the model might be limited by using the softmax loss. Furthermore, the aforementioned model did not take into account inter-class and intra-class distances jointly.

Although most face recognition applications still suffer from the deficiency of difficult facial images with partial occlusions. An intuitive solution to this problem is that more occluded facial images should be included into the training process of the CNN framework. Our goal is to urge our network to learn the face features in difficult cases. Due to loyalty to the different methods of data augmentation, such as flipping and histograms, we add masked parts of faces. More specifically, square patches are cropped from the original image with a random size, thus getting a general framework for the association of face occluded faces and partial faces with a full face. Besides, despite the presence of complex situations, like the occluded faces, in order to study the model of stabilization, we achieve a significant performance in terms of recognition rate (99.2% on the LFW dataset and 96.83% on the YTF dataset). This shows the ability of the CNN model for imperfect facial data analysis without greatly reducing the recognition rate. Furthermore, we adopt the fusion of the softmax loss and the center loss as supervision signals, which help improve the performance and conduct the final classification, unlike studies with softmax only which was not specifically designed for complex samples.

5.4 Performance comparison and evaluation

The comparison with the most recent state of the art on the two LFW and YTF datasets is given in this section.

Firstly, Table 5 depicts the face verification rates on LFW. Our model achieves 99.2% accuracy. The result of our model outperforms the performance of DeepID2 [38], VGG [21], DeepFace [6], DeepID2+ [39], WebFace [40] also Joint-Alex [26], Joint-Res [26] and Light CNN9 [25]. However, our best model is close to the accuracy rate of FaceNet [18] by about 0.66% because FaceNet is trained on a large database that contains 200 million photographs of eight million persons. In addition, having a clearly significant capability, FaceNet adopts the triplet loss function.

The majority of the aforementioned models do not take into account inter-class and intra-class distances jointly. Obviously, the ArcFace method [41], as illustrated in Table 5, achieves up to 99.83% accuracy on the LFW dataset when training with the improved ResNet100 model. The ArcFace method has a clear geometric interpretation due to the exact correspondence to the geodesic distance on the hypersphere. This method achieved an important recognition rate compared to our work. Nevertheless, when there are millions of identities in the training data, ArcFace causes significant training difficulties, e.g., excessive GPU memory consumption and massive computational cost, even at a prohibitive level. Therefore, the authors in [41] used eight GPU cards (four NVIDIA Tesla P40 (24 GB) GPUs) on the training data to achieve high performances. These performances remain in an environment which is not rich by hard and noisy samples. However, in our work, we used DA to make difficult situations on the training data, and our performances do not decrease and we achieve a significant recognition rate of 99.2% on LFW dataset while considering inter-class and intra-class distances jointly.

We also evaluate our model on YTF to further prove its generalization. The results are reported in Table 6. It can

Table 5 Comparison with existing state of the art on LFW dataset in terms of accuracy (%)

Method	Networks	Accuracy (%)
VGG [21]	4	98.37
DeepID2 [38]	100	97.45 ± 0.26
DeepFace [6]	3	97.15 ± 0.27
WebFace [40]	–	97.73 ± 0.31
DeepID2+ [39]	25	98.97 ± 0.25
FaceNet [18]	1	99.63 ± 9
Joint-Alex [26]	1	98.03 ± 0.23
Joint-Res [26]	1	98.70 ± 0.16
Light CNN 9 [25]	1	98.13
ArcFace (ResNet100) [41]	1	99.83
Our model	1 1	99.2 ± 0.04

Table 6 Comparison with existing state of the art on YTF dataset in terms of accuracy (%)

Method	Networks	Accuracy (%)
VGG [21]	4	97.30
DeepID2 + [39]	25	93.20 ± 0.2
DeepFace [6]	3	91.40 ± 1.1
WebFace [40]		92.24 ± 1.28
Joint-Alex [26]	1	92.32 ± 0.40
Joint-Res [26]	1	93.12 ± 0.43
ArcFace [41]	1	98.02
CosFace [43]	1	97.6
SeqFace [42]	1	98.12
Our model	1	96.63

be observed that the verification accuracy of our model outperforms DeepID2+ [39], DeepFace [6] WebFace [40], Joint-Res [26] and Joint-Alex [26] models. CosFace [42] and SeqFace [43] attain important performances on the YTF dataset. However, their performances largely depend on these methods, which require a significant number of iteration steps during training. Also, these methods did not utilized complex environment as our work.

6 Conclusion

In this paper, we have developed a CNN framework to learn a robust face verification on an uncontrolled environment. We have used aggressive DA including randomly perturbing information and complicated conditions for the appearance of faces. One of the key ideas has been to use the adaptive fusing strategy of softmax loss and center loss, which is helpful to improve the performance and to make the model more efficient and flexible. The experimental results on LFW and YTF datasets in the unconstrained face recognition demonstrate that the features extracted by the inception model significantly improve the face recognition performance in verification scenario. Finally, we achieve 99.2% on the LFW and 96.63% on the YTF with the CNN model only, which demonstrate the effectiveness of our method.

Compliance with ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

1. Xu, Y., Zhu, Q., Fan, Z., Zhang, D., Mi, J., Lai, Z.: Using the idea of the sparse representation to perform coarse to-fine face recognition. Inf. Sci. **238**, 138–148 (2013)
2. He, K., Zhang, X., Ren, S., Sun, J.: Deep residual learning for image recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 770–778 (2016)
3. Redmon, J., Divvala, S., Girshick, R., Farhadi, A.: You only look once: unified, real-time object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 779–788 (2016)
4. Lei, Z., Chu, R., He, R., Liao, S., Li, S. Z.: Face recognition by discriminant analysis with Gabor tensor representation. In: International Conference on Biometrics, pp. 87–95. Springer, Berlin (2007)
5. Sun, Y., Chen, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, pp. 1988–1996 (2014)
6. Taigman, Y., Yang, M., Ranzato, M.A., Wolf, L.: Deepface: closing the gap to human-level performance in face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1701–1708 (2014)
7. Yi, D., Lei, Z., Liao, S., Li, S.Z.: Learning face representation from scratch. arXiv:1411.7923 (2014)
8. Choi, J.Y.: Spatial pyramid face feature representation and weighted dissimilarity matching for improved face recognition. Vis. Comput. **34**(11), 1535–1549 (2018)
9. Szegedy, C., Liu, W., Jia, Y., et al.: Going deeper with convolutions. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1–9 (2015)
10. LeCun, Y., Boser, B., Denker, J.S., Henderson, D., Howard, R.E., Hubbard, W., Jackel, L.D.: Backpropagation applied to handwritten zip code recognition. Neural Comput. **1**(4), 541–551 (1989)
11. Yaeger, L., Lyon, R., Webb, B.: Effective training of a neural network character classifier for word recognition. In: Advances in Neural Information Processing Systems, pp. 807–813 (1996)
12. Krizhevsky, A., Sutskever, I., Hinton, G.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
13. Faiedh, H., Hamdi, S., Bouguezzi, S., Farhat, W., Souani, C.: Architectural exploration of multilayer perceptron models for on-chip and real-time road sign classification. Pro. Inst. Mech. Eng. Part I J. Syst. Control Eng. **232**(6), 772–783 (2018)
14. Farhat, W., Sghaier, S., Faiedh, H., Souani, C.: Design of efficient embedded system for road sign recognition. J. Ambient Intell. Humanized Comput. **10**, 1–17 (2018)
15. Fredj, H.B., Ltaif, M., Ammar, A., Souani, C.: Parallel implementation of Sobel filter using CUDA. In: International Conference on Control Automation and Diagnosis (ICCAD), pp. 209–212 (2017)
16. Wang, B., Chen, S., Wang, J., Hu, X.: Residual feature pyramid networks for salient object detection. Vis. Comput. **35**, 1–12 (2019). https://doi.org/10.1007/s00371-019-01779-3
17. Xi, P., Guan, H., Shu, C., Borgeat, L., Goubran, R.: An integrated approach for medical abnormality detection using deep patch convolutional neural networks. Vis. Comput. **35**, 1–14 (2019). https://doi.org/10.1007/s00371-019-01775-7
18. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: a unified embedding for face recognition and clustering. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 815–823 (2015)
19. Sun, Y., Wang, X., Tang, X.: Deep learning face representation from predicting 10,000 classes. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1891–1898 (2014)
20. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems, pp. 1097–1105 (2012)
21. Chatfield, K., Simonyan, K., Vedaldi, A., Zisserman, A.: Return of the devil in the details: delving deep into convolutional nets, 2014. arXiv:1405.3531
22. Guo, K., Wu, S., Xu, Y.F.: Face recognition using both visible light image and near-infrared image and a deep network. CAAI Trans. Intell. Technol. **2**(1), 39–47 (2017)
23. An, F., Liu, Z.: Facial expression recognition algorithm based on parameter adaptive initialization of CNN and LSTM. Vis. Comput. **35**, 1–16 (2019). https://doi.org/10.1007/s00371-019-01635-4
24. Lv, J.J., Cheng, C., Tian, G.D., Zhou, X.D., Zhou, X.: Landmark perturbation-based data augmentation for unconstrained face recognition. Sig. Process. Image Commun. **47**, 465–475 (2016)
25. Wu, X., He, R., Sun, Z., Tan, T.: A light cnn for deep face representation with noisy labels. IEEE Trans. Inf. Forensics Secur. **13**(11), 2884–2896 (2018)
26. Zhang, Y., Shang, K., Wang, J., Li, N., Zhang, M.M.: Patch strategy for deep face recognition. IET Image Proc. **12**(5), 819–825 (2018)
27. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: European

Conference on Computer Vision, pp. 499–515. Springer, Cham (2016)
28. Wen, G., Chen, H., Cai, D., He, X.: Improving face recognition with domain adaptation. Neurocomputing **287**, 45–51 (2018)
29. Devries, T., Biswaranjan, K., Taylor, G.W.: Multi-task learning of facial landmarks and expression. In: Canadian Conference on Computer and Robot Vision, Montreal, QC, pp. 98–103 (2014)
30. Wang, X., Wang, K., Lian, S.: A survey on face data augmentation. arXiv:1904.11685 (2019)
31. Naseem, I., Togneri, R., Bennamoun, M.: Linear regression for face recognition. IEEE Trans. Pattern Anal. Mach. Intell. **32**(11), 2106–2112 (2010)
32. Qi, C., Su, F.: Contrastive-center loss for deep neural networks. In: IEEE International Conference on Image Processing (ICIP) pp. 2851–2855 (2017)
33. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process. Lett. **23**(10), 1499–1503 (2016)
34. Deng, J., Guo, J., Zhou, Y., Yu, J., Kotsia, I., Zafeiriou, S.: RetinaFace: single-stage dense face localisation in the wild. arXiv :1905.00641 (2019)
35. Masi, I., Trần, A.T., Hassner, T., Leksut, J.T., Medioni, G.: Do we really need to collect millions of faces for effective face recognition? In: European Conference on Computer Vision, pp. 579–596. Springer, Cham (2016)
36. Leng, B., Yu, K., Jingyan, Q.I.N.: Data augmentation for unbalanced face recognition training sets. Neurocomputing **235**, 10–14 (2017)
37. Lv, J.J., Shao, X.H., Huang, J.S., Zhou, X.D., Zhou, X.: Data augmentation for face recognition. Neurocomputing **230**, 184–196 (2017)
38. Sun, Y., Wang, X., Tang, X.: Deep learning face representation by joint identification-verification. In: Advances in Neural Information Processing Systems, Montreal, Canada, pp. 1988–1996 (2014)
39. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. In: IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, pp. 2892–2900 (2015)
40. Yi, D., Lei, Z., Liao, S., et al.: Learning face representation from scratch. arXiv:1411.7923 (2014)
41. Deng, J., Guo, J., Xue, N., Zafeiriou, S.: Arcface: additive angular margin loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4690–4699 (2019)
42. Hu, W., Huang, Y., Zhang, F., Li, R., Li, W., Yuan, G.: SeqFace: make full use of sequence information for face recognition. arXiv :1803.06524 (2018)
43. Wang, H., Wang, Y., Zhou, Z., Ji, X., Gong, D., Zhou, J., Li, Z., Liu, W. (2018). Cosface: large margin cosine loss for deep face recognition. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5265–5274

Hana Ben Fredj got her fundamental license and MS degree in Microelectronics from the Higher Institute of Informatics and Mathematics of Monastir, Tunisia, in 2011 and 2014, respectively. Currently, she is preparing her Ph.D. degree in Electronics and Microelectronics in the Faculty of Sciences of Monastir. Her main research includes processing image, recognition pattern, parallel architecture and graphics processor.

Safa Bouguezzi got her fundamental license and MS degree in Microelectronics from the Higher Institute of Informatics and Mathematics of Monastir, Tunisia, in 2013. She got her MS degree in Microelectronics from Higher Institute of Applied Sciences and Technology Sousse, Tunisia. Currently, she is preparing her PhD degree in Electronics and Microelectronics in the Faculty of Sciences of Monastir. Her main research includes embedded system, processing image, recognition pattern, parallel architecture.

Chokri Souani is Professor in Electronics and Microelectronics, at Higher Institute of Applied Sciences and Technology Sousse, Tunisia. He is currently team leader in the Microelectronics and Instrumentation Laboratory µEI (LR13ES12). His research interests include software-defined system, SDR, SD-SoCSoC, MPSoC, embedded system, computer vision, big data, IoT, smart city, communicant vehicle and ITS, small satellite and applications.