applied sciences

MDPI

Article

Boosting Face Recognition under Drastic Views Using a Pose AutoAugment Manner

Wanshun Gao [1], Xi Zhao [2,3] and Jianhua Zou [1,*]

[1] School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an 710049, China; g-wanshun@stu.xjtu.edu.cn

[2] School of Management, Xi'an Jiaotong University, Xi'an 710049, China; zhaoxi1@gmail.com

[3] The Key Lab of the Ministry of Education for Process Control & Efficiency Engineering, Xi'an 710049, China

* Correspondence: jhzou@sei.xjtu.edu.cn

Received: 14 October 2020; Accepted: 28 November 2020; Published: 15 December 2020

check for updates

Abstract: Face recognition under drastic pose drops rapidly due to the limited samples during the model training. In this paper, we propose a pose-autoaugment face recognition framework (PAFR) based on the training of a Convolutional Neural Network (CNN) with multi-view face augmentation. The proposed framework consists of three parts: face augmentation, CNN training, and face matching. The face augmentation part is composed of pose autoaugment and background appending for increasing the pose variations of each subject. In the second part, we train a CNN model with the generated facial images to enhance the pose-invariant feature extraction. In the third part, we concatenate the feature vectors of each face and its horizontally flipped face from the trained CNN model to obtain a robust feature. The correlation score between the two faces is computed by the cosine similarity of their robust features. Comparable experiments are demonstrated on Bosphorus and CASIA-3D databases.

Keywords: pose autoaugment; face recognition; convolutional neural network; face augmentation

1. Introduction

Face recognition is one of the most studied topics in computer vision. It has been widely applied in security, pedestrian identification, and other fields. Traditional face recognition methods are mainly based on handcrafted features, such as High-Dimensional Local Binary Patterns (HD-LBP) [1], Fisher Vector (FV) descriptors [2], and Multi-Directional Multi-Level Dual-Cross Patterns (MDML-DCPs) [3]. However, handcrafted features are not robust. As Convolutional Neural Network (CNN) provides a better solution for this problem, many face recognition approaches based on CNN have emerged. Incorporating deeper networks and large training sets, CNN-based approaches [4,5] have transcended human performance on the Labeled Faces in the Wild (LFW) [6] dataset. Despite the dominance of existing CNN-based face recognition approaches in feature extraction, there is still a challenge that the recognition accuracy of profile faces (one eye is self-occluded) drops rapidly [7,8]. The main reason is that the training sets, such as CASIA WebFaces, which is crawled from the Internet, are not evenly distributed on the head pose [9,10]. The insufficient intra-class variations (differences of poses in a subject) make the recognition model less sensitive to profile faces [9].

In response to the challenge of profile face recognition, two novel types of face recognition methods are proposed, i.e., facial image normalization [11,12] and face augmentation [9,13,14], both of which normally require 3D face models to ease the difficulty. Facial image normalization transforms multiple views into a uniform one. In [11], the 3D-2D system (UR2D) maps the facial image to UV space using 3D facial data, which achieves better performance than using a reconstructed 3D shape. The landmark detection and pose estimation are essential for UR2D, but may be challenging and time-consuming when transforming drastic face poses to the UV space. By contrast, face augmentation transforms one view into multiple views. Recent studies [9,13,15] map 2D faces to generic 3D shapes or reconstructed 3D shapes, but render only limited views for each identity. The sparse sampling of views may reduce recognition accuracy [16].

In order to tackle the issue of sparse sampling and enhance the recognition accuracy under drastic pose, we propose an enhanced face recognition framework based on CNN training with a random-view face augmentation method. The proposed framework consists of three parts: face augmentation, CNN training, and face matching. In the first part, we aim at increasing intra-class variations of views. With the aid of a 3D graphic engine, numerous views, especially drastic pose views, are generated by face scans. Pose AutoAugment is proposed to find the best distribution of facial views. As these views have no backgrounds, the facial contour may change drastically. In order to avoid the information of facial contours being learned by the following CNN training, we randomly append a background behind each view. In the second part, we train a CNN model with views that are generated in the first part, to enhance the pose-invariant feature extraction. The CNN architecture is adapted from the classical research [17] on face recognition. In the third part, we concatenate the feature vectors of each face from the trained CNN model. The correlation score between two faces is computed by the cosine similarity of their feature vectors. By computing the scores of a facial image under drastic pose to all images corresponding to different subjects, we adopt the highest score to identify an unknown subject in the probe from registered subjects in the gallery. Experiments on the Bosphorus database and CASIA-3D database demonstrate a state-of-the-art performance of the proposed framework. Furthermore, the importance of components in the first part, for example, adding backgrounds and face cropping, are also evaluated.

Our contributions are as follows.

i An enhanced face recognition framework is proposed to identify pose-invariant representations of faces under drastic poses. The proposed framework trains a pose-invariant CNN model, and extract identifiable features of drastic pose view for face recognition.

ii A novel face augmentation method, which is composed of pose auto-augment and background appending, is proposed to increase pose variations for each subject.

iii Experiments on Bosphorus and CASIA-3D FaceV1 databases demonstrate state-of-the-art performance for face recognition under drastic poses.

The rest of this paper is organized as follows. Section 2 briefly reviews related works on pose-invariant face recognition and rendering methods. Section 3 illustrates the face augmentation method and the proposed face recognition framework. Experimental results and discussion are demonstrated in Section 4. This paper is concluded in Section 5.

2. Related Work

With the rapid development of CNN, face recognition has made a lot of progress in the past ten years. Although the maturity of face recognition leads the increasing success from research to commercial application, there are still some challenges, such as occlusion [18], age [19], pose [20], and attack [21]. In pose-invariant face recognition, there are two streams: one is image normalization, and the other is face augmentation. Both streams mainly focus on two issues: 3D shape fitting and face generating. Since the 1970s, 3D models have been used for image generating [22–25]. Many studies apply image generating to object detection, object retrieval [26,27], and viewpoint estimation [28], etc. Recently, image generating methods have been employed for face recognition [24], face alignment [29], and 3D face reconstruction [30].

The image normalization methods reconstruct the facial image to 3D face and generate the same views from 3D faces for comparison. Georghiades et al. [31] employed a reconstructed surface to render nine views, then the test image matched these views in a linear subspace. However, the poses of the rendered face are no more than 24°. Wang et al. [32] reconstructed a 3D face by fitting a 2D facial image and generated multi-view virtual faces. A Gabor feature was extracted from both virtual faces and test faces to identify the same person. Prabhu et al. [33] and Moeini et al. [34] also generated multi-view virtual faces from a reconstructed 3D face, further estimating the viewpoint of the test face for comparing the test face with virtual faces under similar view. Dou et al. [35] reconstructed accurate 3D shapes to transform the pose variant of images to a uniform facial space.

The face augmentation methods are proposed to fit facial images to 3D shapes and generate multi-view facial images to learn pose-invariant features for comparison. Rather than concentrating on accurate 3D reconstruction, Hassner et al. rendered the frontal face with a generic 3D face shape [36]. Then 10 generic shapes were employed to render new facial views [9,13]. Another idea of 3D shape fitting is fitting a facial image to its real 3D shape, which was proposed by Kakadiaris et al. [11]. In terms of face generating, Hassner et al. [36] only generated a frontal face for each identity. Vasilescu et al. [24,37] rendered 15 face images from -35° viewpoint to $+35^\circ$ viewpoint in 5° steps, including six viewpoints for training and nine viewpoints for testing. Hassner et al. also generated five views ($0^\circ, \pm45^\circ, \pm75^\circ$) to train a pose-invariant CNN model [9,13]. Crispell et al. [15] further developed the idea of face generating [9], and generated five views randomly for each identity.

Both image normalization and face augmentation methods generate limited views for face recognition. However, the sparse sampling of views may reduce the recognition accuracy [16]. Dou et al. [30] generate numerous views, but to design an end-end 3D reconstruction system. We adapt the idea of image generation [30] to our research. Different from [30], we generate numerous views with searched distribution to train a CNN model aimed at the performance improvement of face recognition under drastic poses. Furthermore, our face augmentation method is compatible with any type of 3D face, and we use face scans in this paper, due to its accurate 3D shape and simplicity of texture mapping to 3D shape via camera calibration.

To our best knowledge, we are the first to generate arbitrary views that are more accurate and realistic for face training and recognition.

3. Proposed Methods

We propose a novel framework (PAFR) for face recognition under drastic pose. As shown in Figure 1, the proposed framework consists of three parts: face augmentation, CNN training using generated faces, and face matching. In the following section, we describe these three parts in detail.

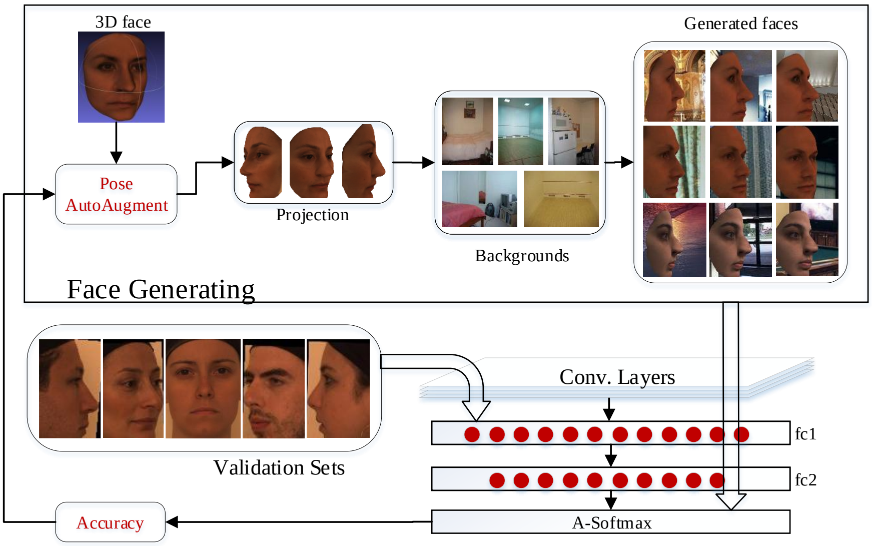

Figure 1. The framework of PAFR. First, we preprocess the face scan to obtain a smoothed face, then randomly render 2D faces by Pose AutoAugment. 2D synthetic faces are generated as training samples after projection, cropping, background appending. Third, we train an arbitrary-pose face recognition model. The controller model is updated by the validation accuracy.

3.1. Face Augmentation

To increase intra-class pose variations for each subject, we employ a 3D graphic engine to generate multi-views from a 3D face. Both types of the 3D face (reconstructed 3D face and face scan) are compatible with our face augmentation method. In this paper, we use face scans due to their accurate 3D shape and simplicity of texture mapping to a 3D shape via camera calibration. The face augmentation pipeline for face scans consists of four components: preprocessing, pose autoaugment, face cropping, and random background appending.

Preprocessing is necessary to obtain a smoothed face scan since a raw face scan contains lots of noise, irregular edges, and hollows. Irregular edges and hollows mainly appear in the eyes, hair, and ears, due to their reflective properties and self-occlusion. Surface noise is substantially affected by the precision of scanners. To remove noise, irregular edges, and hollows, preprocessing involves cropping the facial region, filling hollows, and denoising. First, the position of the nose tip is located. The points in which the distance from the nose tip exceeds a threshold are removed. Then, we use the bilinear interpolation algorithm to fill missing points. Lastly, we employ the Laplacian algorithm to obtain a smoothed face scan.

In terms of pose autoaugment, we simulate a camera in a 3D context to project a face scan to 2D faces with a 3D graphic engine. Millions of views from a single 3D face can be generated by exploring the extrinsic parameters of the camera: azimuth R_a, elevation R_e, and in-plane R_i rotation. Although a pose-invariant face recognition model performs better with more views, it may induce a heavy load if we use too many views in the following tasks, such as cropping, background adding, and training. To keep a balance between the variety of poses and the resource consumption, we generate a repository of enough poses with searched distribution, and randomly sample poses from the repository for each subject.

To be specific, a number N and distribution parameters W_P are set to generate enough extrinsic parameters $\{R_a, R_e, R_i\}$. Then a group of parameters for each subject is randomly sampled from these $\{R_a, R_e, R_i\}$. Given N and W_P, W_X is first generated obeying the uniform distribution of $(0, W_{total})$, where $W_P = \{w_1, w_2, ...w_i, ..., w_m\}$, $W_X = \{W_{x_1}, W_{x_2}, ..., W_{x_N}\}$ and $W_{total} = \sum_{i=1}^{m} w_i$. X is an integer vector $\{x_1, x_2, ..., x_j, ..., x_n\}$, and x_j is obtained by computing

$$W_{x_j} \leq \sum_{i=1}^{x_j} w_i \quad and \quad W_{x_j} \geq \sum_{i=1}^{x_j-1} w_i. \tag{1}$$

To those rotation parameters R for a subject,

$$R = sX + B, \tag{2}$$

where s is a scale value. $B = \{b_1, b_2, ..., b_n\}$, obeys the uniform distribution of $(-\theta, \theta)$. θ and s are set manually.

By changing W_P and N, we can obtain the distribution and as many rotation parameters as we want. To improve the recognition performance on profile faces automatically, we tune the distribution W_P and N to find the proper number of profiles and the near-frontal faces using Bayesian Optimization [38].

We crop the generated faces by reserving the same aspect ratio but do not align them for the following two reasons. First, as suggested by the literature on VGG-face training [39], face recognition achieves better performance when training faces are not aligned. Furthermore, when the pose is larger than 60 degrees, the aligned face will be seriously distorted as a result of the occluded eye corner and mouth corner.

The background of each generated face is transparent, leading to high contrast on the facial contour. To prevent the CNN classifier from overfitting unrealistic contour patterns, we synthesize the background in a flexible manner by randomly appending a scene image as the background. The alpha channel is employed to combine the generated face and background.

3.2. Face Recognition

3.2.1. CNN Training

Recent face recognition models [4,5,17] achieve state-of-the-art accuracy on YTF, LFW, and Megaface. As we enrich large-pose faces as the training set, better performance is expected on faces under drastic pose than existing CNN-based methods.

The training set consists of generated faces from frontal scans of 3D databases, such as Bosphorus and CASIA-3D FaceV1 in this paper. The CNN architecture is adapted from SphereFace CNN [17] with 20 layers, which is trained on CAISA-WebFace. We freeze the parameters of all convolution layers, and fine-tune the parameters of the fully connected layers for mapping the feature $\boldsymbol{f}_i$ from the last convolution layer to its identity label $\boldsymbol{C}_i$. The minimization of loss function L (Equation (3)) is to maximize $cos(m\theta_{C_i,f_i})$ and minimize $cos(\theta_{j,f_i})$.

$$L = \frac{1}{N}\sum_i -log\left(\frac{e^{\|\boldsymbol{f}_i\|cos(m\theta_{C_i,f_i})}}{e^{\|\boldsymbol{f}_i\|cos(m\theta_{C_i,f_i})} + \sum_{j\neq C_i} e^{\|\boldsymbol{f}_i\|cos(\theta_{j,f_i})}}\right) \tag{3}$$

where θ_{C_i,f_i} is the angle between $\boldsymbol{f}_i$ and its identity label C_i. θ_{j,f_i} is the angle between $\boldsymbol{f}_i$ and other identity label j. $\theta \in [0,\pi]$. m is the angular margin. N is the number of generated images. When maximizing $cos(m\theta_{C_i,f_i})$, the angle θ_{C_i,f_i} will be minimized to zero. When minimizing $cos(\theta_{j,f_i})$, the angle θ_{j,f_i} will be maximized to π.

3.2.2. Face Matching

In the testing phase, we drop the last fully connected layer of trained CNN, which indicates a label of face identity, and adopt the penultimate layer that keeps high-level information as our feature representation. The testing set consists of photos. These photos are not used to reconstruct 3D faces and generate new views for face matching, since landmark detection and pose estimation is essential for 3D reconstruction, which may not be accurate for drastic face pose. These photos are matched based on the proposed framework. For a photo $\boldsymbol{p}$, we extract a feature vector $\boldsymbol{f}_p$ and its horizontally flipped feature $\boldsymbol{f}_{flip}$ from the penultimate layer of trained CNN. A robust face representation $\boldsymbol{r}_p$ (Equation (4)) is achieved by concatenating these feature vectors.

$$\boldsymbol{r}_p = [\boldsymbol{f}, \boldsymbol{f}_{flip}]. \tag{4}$$

For a face pair denoted as ($\boldsymbol{p}_1$, $\boldsymbol{p}_2$), the score s (Equation (5)) is computed by the cosine similarity of their robust representation.

$$s(\boldsymbol{p}_1, \boldsymbol{p}_2) = \frac{\boldsymbol{r}_1^T\boldsymbol{r}_2}{||\boldsymbol{r}_1||\ ||\boldsymbol{r}_2||}. \tag{5}$$

Before face matching, near-frontal faces (yaw rotations in ±30°) in the testing set are aligned using five landmarks (two eyes, two mouth corners, and nose) while the other faces are aligned using the visible eye and the tip of the nose. MTCNNv1 [40] is employed for detecting and aligning the testing faces. However, MTCNNv1 is so not efficient on profile faces. The failure cases in face detection are manually aligned.

4. Experiments

4.1. Datasets

To evaluate the performance of PAFR on faces under drastic poses, databases contain 2D arbitrary poses and frontal face scans are needed. CASIA-3D FaceV1 and Bosphorus were employed for our experiments. Bosphorus contains 105 individuals and 4652 scans in total. There are 60 men and 45 women. Most of them are between 25 and 35. Each one has no less than 31 scans, and no more than 54 scans. Facial poses contain seven yaw rotations ranging from −90° to 90°, four pitch rotations, and two cross rotations. Each scan contains a point cloud, a facial image, and 22 manually labeled feature points. The facial image can be mapped to 3D space as a face texture by coordinate mapping. CASIA-3D FaceV1 contains 4624 scans of 123 individuals, 37 or 38 scans per person, covering the variation of the pose, emotion, and illumination. Each scan has a 3D face with texture and a facial image. Compared with Bosphorus, the faces in CASIA-3D FaceV1 are darker, since the light source in CASIA-3D FaceV1 is incandescent light. In Table 1, a summary of the datasets is presented.

Table 1. A brief summary of the datasets in our experiments.

Datasets	Persons	Training Set (3D Frontal Scans)	Testing Set (2D Facial Images)	Testing Set (Facial Pose)
Bosphorus	105	105	1365	7 yaw, 4 pitch, and 2 cross rotations
CASIA-3D Face V1	122	122	976	6 yaw, and 2 pitch rotations.

For experiments on Bosphorus, 105 frontal scans, 1355 facial images under 13 poses are employed including 101 facial images under −90° and 100 facial images under 90°. The frontal scans are employed for generating faces as the training set, 2D faces under 13 poses as the testing set. For experiments on CASIA-3D FaceV1, the training set consists of faces generated by 122 frontal scans under office light, and the testing set consists of 2D faces under eight poses under the office light. Figure 2 shows the 2D faces in Bosphorus and CASIA-3D FaceV1.

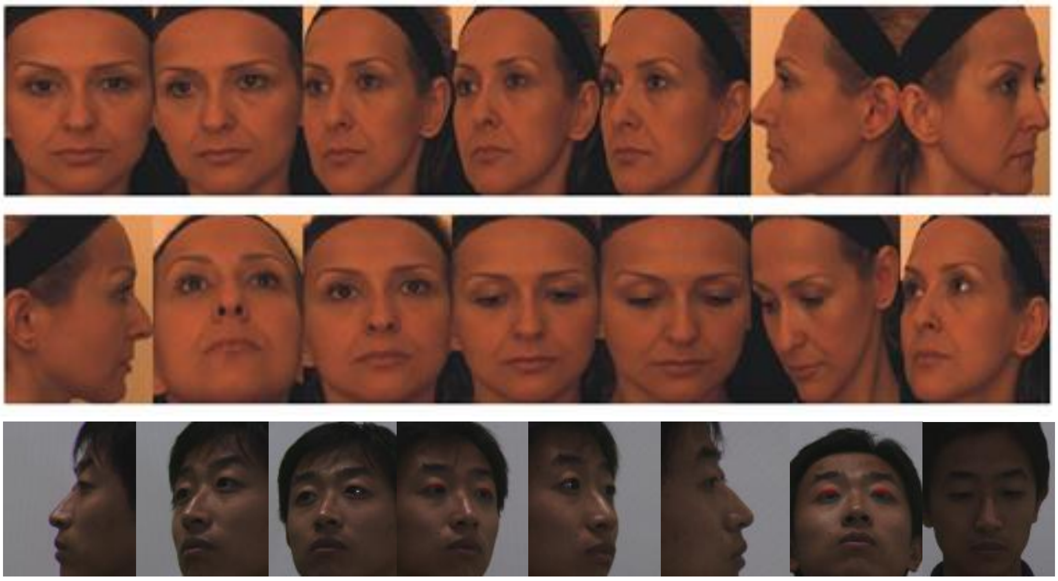

Figure 2. 2D faces in Bosphorus and CASIA-3D FaceV1. In Bosphorus, facial poses of each identity have 7 yaw rotations, ranging from −90° to 90°, 4 pitch rotations, and 2 cross rotations. In CASIA-3D FaceV1, facial poses of each identity have 6 yaw rotations, ranging from −90° to 90°, and 2 pitch rotations.

For face augmentation, we adjust the camera settings, including the intrinsic and extrinsic parameters. For intrinsic parameters, the focal length is 35 and aspect ratio is 1.0. For extrinsic parameters, the rotation angles in azimuth, elevation and in-plane are in the range of $[-90°, 90°]$, $[-30°, 30°]$ and $[-5°, 5°]$, respectively. A 3D graphic engine, Blender, is employed, for its open-source and python-support. The backgrounds on generated faces are from the SUN397 database [41]. Figure 3 shows the faces generated by face scans in Bosphorus. Each scan randomly generates 512 facial images, from which we selected 15 images that represent multi-view faces under yaw rotations (azimuth rotations to the camera) ranging from −90° to 90°. It can be seen that the generated faces are under arbitrary pose with high fidelity.

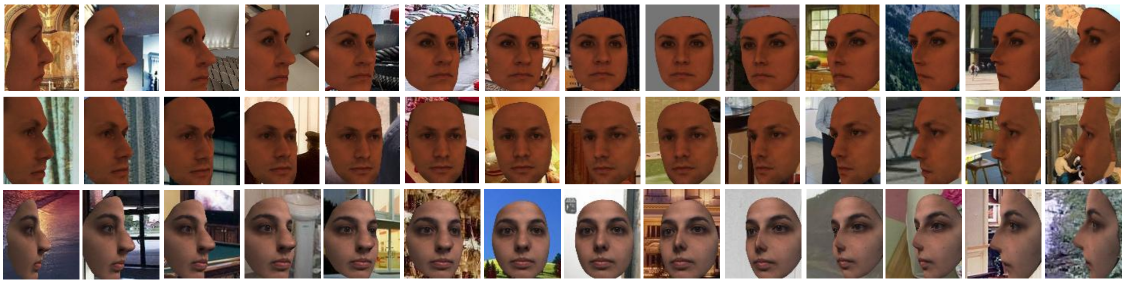

Figure 3. Generated faces on Bosphorus. Fifteen views of each identity are selected from 512 random views that represent multi-view faces ranging from −90° to 90°.

4.2. *Evaluate the Components in Face Augmentation*

To evaluate the impact of each component in the face augmentation, training faces are generated by the proposed augmentation method except for one or two components. After CNN training is finished, the class of the max probability out of the Softmax classifier is the predicted class (named as rank-1 accuracy). We compute the rank-1 accuracy for 2D faces from Bosphorus.

The baseline is illustrated by computing the accuracy on Bosphorus. First, we generate faces without cropping and background appending. The accuracy is reported as ′cbN′. Second, we combine backgrounds randomly on generated faces of cbN. The accuracy is shown as ′cBN′. Third, the generated faces of the cbN are cropped as ′CbN′. Forth, we combine backgrounds randomly on generated faces of CbN. The accuracy is reported as ′CBN′. Table 2 shows these reports.

Table 2 depicts the benefits of the combination of components. When cropping on faces under ±90° pose, the performance is increased by more than 8% without a background, and by more than 20.79% with a background appending. When not cropping on faces under ±90° pose, the performance is increased by more than 1% without a background, and by more than 2.66% after cropping. Experimental results justify the effectiveness of the cropping component and the background appending component.

Table 2. Accuracy of the combination of components. ′C′ and ′c′ represents face cropping and without cropping, ′B′ and ′b′ represents face rendering with a background without the background, ′N′ represents rendering 512 faces each scan, L45 and R45 represent facial images with −45 and 45 degrees on the horizontal plane, L90 and R90 represent facial images with −90 and 90 degrees on the horizontal plane.

Pose	L45	R45	L90	R90
None	100	100	71	66.34
cbN	100	100	72	67.5
cBN	100	100	75	69
CbN	100	100	79	71.29
CBN	100	100	94	87.13

4.3. *Evaluate the Face Recognition under Pose Variations*

To validate the effectiveness of the proposed framework under arbitrary poses, we calculate the accuracy of face identification on Bosphorus and CASIA-3D FaceV1. We set nine groups of gallery and probe. To all nine groups, the probes are the same, and consist of faces under poses ranging from left 90° to right 90°. For each gallery, the faces are under the same views. Tables 3 and 4 demonstrate the effectiveness and robustness of PAFR.

Table 3. Rank-1 identification accuracies on Bosphorus under pose rotations. R10 represents facial images under yaw rotations of 10 degrees, similar to R20, R30, R45, R90. L45 represents yaw rotations of −45 degrees, similar to L90. CR represents yaw rotations of 45 degrees and pitch rotations of 20 degrees. SU and SD represent pitch rotations of slight upwards and slight downwards. Up and Down represent pitch rotations of upward and downward.

Probe Gallery	L45-R45	SU&SD	Up	Down	CR	L90	R90
Front Face	100	100	100	100	100	94	87.13
R10	100	100	99	100	100	91	89.11
R20	100	100	99	100	100	88	87.13
R30	100	100	98.10	100	100	91	86.14
R45	100	100	98.10	100	100	93	88.12
L45	100	100	97.14	100	100	96	84.16
R90	83.05	82.38	64.76	77.88	78.57	78	N/A
L90	85.14	87.62	72.38	77.88	80	N/A	74.26

Table 3 describes the performance of the proposed framework on Bosphorus. It shows little difference in face identification when poses of faces in the gallery are from left 45° to right 45°, respectively. When the gallery consists of frontal faces, the recognition accuracy is 94% for the probe, which consists of faces under left 90° rotations. However, the recognition accuracy drops significantly when the gallery only consists of faces under ±90°.

Table 4. Rank-1 identification accuracies on CASIA-3D FaceV1 under pose rotations. R30 represents facial images under yaw rotations of 30 degrees, similar to R60 and R90. L30 represents yaw rotations of −30 degrees, similar to L60 and L90. Up and Down represent pitch rotations of upward and downward.

Probe Gallery	L30-R30	L60	R60	L90	R90	Up	Down
Front Face	100	100	100	93.33	92.62	100	100
L30	100	100	100	95.90	93.33	100	100
R30	99.73	100	99.18	95	91.80	100	99.18
L60	99.72	-	100	95.90	94.17	99.18	99.18
R60	99.45	99.18	-	91.80	95	96.72	98.36
L90	89.88	92.62	86.89	-	87.50	87.70	84.43
R90	90.16	87.70	92.62	87.70	-	88.52	84.43
Up	100	100	98.36	91.80	86.67	-	98.36
Down	99.45	98.36	99.18	90.16	90	97.54	-

Table 4 describes the similar results on CASIA-3D FaceV1 as in Table 3. The performance on CASIA-3D FaceV1 is better than the performance on Bosphorus. Both results have validated the effectiveness of PAFR under arbitrary poses, especially drastic pose.

4.4. Comparison with State-of-the-Art

In Table 5, we compare the performance of PAFR with recent researches on Bosphorus. PGM [42], PGDP [43], and Liang et al. [44] demonstrated the state-of-the-art performance in 3D face recognition. FLM + GT [45] and Sang et al. [46] demonstrated the performances of image normalization methods. It can be observed that the proposed framework outperforms 3D face recognition methods and the image normalization method at all poses, except for the accuracy on faces under right 90°, which is lower than [46]. When examining the failed cases under R90°, we found that most of these failed cases are not well-aligned. Though we employ MTCNNv1 for aligning the testing faces before face matching, MTCNNv1 is not efficient on profile faces. Since testing faces under L90° are better aligned than those under R90°, the accuracy of L90° is higher.

Table 5. Comparison with state-of-the-art.

Pose Methods	R10-30	L45	R45	L90	R90	PU	PD	CR	Average
PGM [42]	87.1	38.1	39.0	N/A	N/A	79.0	69.5	49.1	69.1
PGDP [43]	89.2	37.1	36.2	N/A	N/A	91.4	70.0	50.0	71.4
FLM+GT [45]	N/A	99.0	98.8	83.2	83.0	N/A	N/A	95.7	92.0
Sang et al. [46]	99.8	98.1	97.7	92.3	91.4	93.0	92.7	90.1	93.4
Liang et al. [44]	98.4	93.3		N/A	N/A	99.5		94.8	97.2
PAFR	100	100	100	94.0	87.1	100	100	100	98.1

5. Conclusions

We propose a pose-autoaugment face recognition framework, which is the first to generate arbitrary views that are more accurate and realistic for face training and recognition. The proposed framework trains a pose-invariant CNN model, and extracts identifiable features of drastic pose view for face recognition. The proposed face augmentation method increases the pose variations for each subject. The experiments on Bosphorus show that our work improves the average accuracy over all poses. The experiments also demonstrate the robustness and effectiveness under drastic poses. In the future, we will explore face augmentation methods with 3D faces scanned by low-resolution devices, such as Kinect.

Author Contributions: Conceptualization, W.G.; methodology, W.G.; writing—original draft preparation, W.G.; writing—review and editing, X.Z.; supervision, X.Z. and J.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Acknowledgments: This research was funded by National Natural Science Foundation of China (Grant No. 91746111), The Key Research and Development Program of Shaanxi Province (Grant No. 2020ZDLGY09-08)

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Chen, D.; Cao, X.; Wen, F.; Sun, J. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Portland, OR, USA, 23–28 June 2013; pp. 3025–3032.
2. Simonyan, K.; Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Fisher Vector Faces in the Wild. In Proceedings of the British Machine Vision Conference, Bristol, UK, 9–13 September 2013; pp. 8.1–8.11.
3. Ding, C.; Choi, J.; Tao, D.; Davis, L.S. Multi-directional multi-level dual-cross patterns for robust face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* **2016**, *38*, 518–531. [CrossRef] [PubMed]
4. 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, Boston, MA, USA, 7–12 June 2015; pp. 815–823.
5. Sun, Y.; Wang, X.; Tang, X. Sparsifying Neural Network Connections for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 4856–4864.
6. Huang, G.B.; Ramesh, M.; Berg, T.; Learned-Miller, E. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*; Technical Report 07-49; University of Massachusetts: Amherst, MA, USA, 2007.
7. Zhu, Z.; Luo, P.; Wang, X.; Tang, X. Multi-view perceptron: A deep model for learning face identity and view representations. In Proceedings of the Advances in Neural Information Processing Systems, Montreal, QC, Canada, 8–13 December 2014; pp. 217–225.

8. Yim, J.; Jung, H.; Yoo, B.; Choi, C.; Park, D.; Kim, J. Rotating your face using multi-task deep neural network. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 676–684.
9. Masi, I.; Tran, A.Â.; Hassner, T.; Leksut, J.T.; Medioni, G. Do we really need to collect millions of faces for effective face recognition? In Proceedings of the European Conference on Computer Vision, Amsterdam, The Netherlands, 11–14 October 2016; pp. 579–596.
10. Masi, I.; Chang, F.J.; Choi, J.; Harel, S.; Kim, J.; Kim, K.; Leksut, J.; Rawls, S.; Wu, Y.; Hassner, T.; et al. Learning Pose-Aware Models for Pose-Invariant Face Recognition in the Wild. *IEEE Trans. Pattern Anal. Mach. Intell.* **2018**, *412*, 379–393. [CrossRef] [PubMed]
11. Kakadiaris, I.A.; Toderici, G.; Evangelopoulos, G.; Passalis, G.; Chu, D.; Zhao, X.; Shah, S.K.; Theoharis, T. 3D-2D face recognition with pose and illumination normalization. *Comput. Vis. Image Underst.* **2017**, *154*, 137–151. [CrossRef]
12. Yin, X.; Yu, X.; Sohn, K.; Liu, X.; Chandraker, M. Towards large-pose face frontalization in the wild. In Proceedings of the IEEE International Conference on Computer Vision, Venice, Italy, 22–29 October 2017; pp. 4010–4019.
13. Masi, I.; Hassner, T.; Tran, A.T.; Medioni, G. Rapid synthesis of massive face sets for improved face recognition. In Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Washington, DC, USA, 30 May–3 June 2017; pp. 604–611.
14. Chang, F.J.; Tran, A.T.; Hassner, T.; Masi, I.; Nevatia, R.; Medioni, G. FacePoseNet: Making a case for landmark-free face alignment. In Proceedings of the IEEE International Conference on Computer Vision Workshops, Venice, Italy, 22–29 October 2017; pp. 1599–1608.
15. Crispell, D.; Biris, O.; Crosswhite, N.; Byrne, J.; Mundy, J.L. Dataset Augmentation for Pose and Lighting Invariant Face Recognition. *arXiv* **2017**, arXiv:1704.04326.
16. Kortylewski, A.; Egger, B.; Schneider, A.; Gerig, T.; Forster, A.; Vetter, T. Empirically Analyzing the Effect of Dataset Biases on Deep Face Recognition Systems. *arXiv* **2017**, arXiv:1712.01619.
17. Liu, W.; Wen, Y.; Yu, Z.; Li, M.; Raj, B.; Song, L. SphereFace: Deep Hypersphere Embedding for Face Recognition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 6738–6746.
18. Yuan, X.; Park, I.K. Face De-Occlusion Using 3D Morphable Model and Generative Adversarial Network. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 10061–10070. [CrossRef]
19. Sawant, M.M.; Bhurchandi, K.M. Age invariant face recognition: A survey on facial aging databases, techniques and effect of aging. *Artif. Intell. Rev.* **2019**, *52*, 981–1008. [CrossRef]
20. Ding, C.; Tao, D. A Comprehensive Survey on Pose-Invariant Face Recognition. *ACM Trans. Intell. Syst. Technol.* **2016**, *7*, 37:1–37:42. [CrossRef]
21. Rössler, A.; Cozzolino, D.; Verdoliva, L.; Riess, C.; Thies, J.; Nießner, M. FaceForensics++: Learning to Detect Manipulated Facial Images. In Proceedings of the 2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea, 27 October–2 November 2019; pp. 1–11. [CrossRef]
22. Nevatia, R.; Binford, T.O. Description and recognition of curved objects. *Artif. Intell.* **1977**, *8*, 77–98. [CrossRef]
23. Oliver, N.M.; Rosario, B.; Pentland, A.P. A Bayesian computer vision system for modeling human interactions. *IEEE Trans. Pattern Anal. Mach. Intell.* **2000**, *22*, 831–843. [CrossRef]
24. Vasilescu, M.A.O.; Terzopoulos, D. Multilinear independent components analysis. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Diego, CA, USA, 20–26 June 2005; pp. 547–553.
25. Michels, J.; Saxena, A.; Ng, A.Y. High speed obstacle avoidance using monocular vision and reinforcement learning. In Proceedings of the 22nd International Conference on Machine Learning, Bonn, Germany, 7–11 August 2005; pp. 593–600.
26. Li, Y.; Su, H.; Qi, C.R.; Fish, N.; Cohen-Or, D.; Guibas, L.J. Joint embeddings of shapes and images via CNN image purification. *Acm Trans. Graph.* **2015**, *34*, 234. [CrossRef]
27. Guo, H.; Wang, J.; Gao, Y.; Li, J.; Lu, H. Multi-view 3D object retrieval with deep embedding network. *IEEE Trans. Image Process.* **2016**, *25*, 5526–5537. [CrossRef] [PubMed]

28. Su, H.; Qi, C.R.; Li, Y.; Guibas, L.J. Render for cnn: Viewpoint estimation in images using cnns trained with rendered 3D model views. In Proceedings of the IEEE International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2686–2694.
29. Zhu, X.; Lei, Z.; Liu, X.; Shi, H.; Li, S.Z. Face alignment across large poses: A 3D solution. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 27–30 June 2016; pp. 146–155.
30. Dou, P.; Shah, S.K.; Kakadiaris, I.A. End-to-end 3D face reconstruction with deep neural networks. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 1503–1512.
31. Georghiades, A.S.; Belhumeur, P.N.; Kriegman, D.J. From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. *IEEE Trans. Pattern Anal. Mach. Intell.* **2001**, *23*, 643–660. [CrossRef]
32. Wang, L.; Ding, L.; Ding, X.; Fang, C. 2D face fitting-assisted 3D face reconstruction for pose-robust face recognition. *Soft Comput.* **2011**, *15*, 417–428. [CrossRef]
33. Prabhu, U.; Heo, J.; Savvides, M. Unconstrained pose-invariant face recognition using 3D generic elastic models. *IEEE Trans. Pattern Anal. Mach. Intell.* **2011**, *33*, 1952–1961. [CrossRef]
34. Moeini, A.; Faez, K.; Moeini, H. Unconstrained pose-invariant face recognition by a triplet collaborative dictionary matrix. *Pattern Recognit. Lett.* **2015**, *68*, 83–89. [CrossRef]
35. Dou, P.; Zhang, L.; Wu, Y.; Shah, S.K.; Kakadiaris, I.A. Pose-robust face signature for multi-view face recognition. In Proceedings of the Biometrics Theory, Applications and Systems, Arlington, VA, USA, 8–11 September 2015; pp. 1–8.
36. Hassner, T.; Harel, S.; Paz, E.; Enbar, R. Effective face frontalization in unconstrained images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, MA, USA, 7–12 June 2015; pp. 4295–4304.
37. Vasilescu, M.A.O. Multilinear projection for face recognition via canonical decomposition. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops, San Francisco, CA, USA, 13–18 June 2010; pp. 184–191.
38. Martinez-Cantin, R. BayesOpt: A Bayesian optimization library for nonlinear optimization, experimental design and bandits. *J. Mach. Learn. Res.* **2014**, *15*, 3735–3739.
39. Parkhi, O.M.; Vedaldi, A.; Zisserman, A. Deep Face Recognition. In Proceedings of the British Machine Vision Conference, Swansea, UK, 7–10 September 2015; pp. 41.1–41.12.
40. Zhang, K.; Zhang, Z.; Li, Z.; Qiao, Y. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Process. Lett.* **2016**, *23*, 1499–1503. [CrossRef]
41. Xiao, J.; Hays, J.; Ehinger, K.A.; Oliva, A.; Torralba, A. Sun database: Large-scale scene recognition from abbey to zoo. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, San Francisco, CA, USA, 13–18 June 2010; pp. 3485–3492.
42. Hajati, F.; Raie, A.A.; Gao, Y. 2.5D face recognition using patch geodesic moments. *Pattern Recognit.* **2012**, *45*, 969–982. [CrossRef]
43. Gheisari, S.; Javadi, S.; Kashaninya, A. 3D face recognition using patch geodesic derivative pattern. *Int. J. Smart Electr. Eng.* **2013**, *2*, 127–132.
44. Liang, Y.; Zhang, Y.; Zeng, X.X. Pose-invariant 3D face recognition using half face. *Signal Process. Image Commun.* **2017**, *57*, 84–90. [CrossRef]
45. Moeini, A.; Moeini, H. Real-world and rapid face recognition toward pose and expression variations via feature library matrix. *IEEE Trans. Inf. Forensics Secur.* **2015**, *10*, 969–984. [CrossRef]
46. Sang, G.; Li, J.; Zhao, Q. Pose-invariant face recognition via RGB-D images. *Comput. Intell. Neurosci.* **2016**, *2016*, 3563758. [CrossRef] [PubMed]

Publisher's Note: MDPI stays neutral with regard to jurisdictional claims in published maps and institutional affiliations.