FACE RECOGNITION FOR FISHEYE IMAGES

Yi-Cheng Lo[1], Chiao-Chun Huang[1], Yueh-Feng Tsai[1], I-Chan Lo[2], An-Yeu (Andy) Wu[1], and Homer H. Chen[2]

[1]*Graduate Institute of Electronics Engineering, National Taiwan University, Taipei, Taiwan*
[2]*Graduate Institute of Communication Engineering, National Taiwan University, Taipei, Taiwan*
{alan,jorge,yueh}@access.ee.ntu.edu.tw, {f05942036,andywu,homer}@ntu.edu.tw

2022 IEEE International Conference on Image Processing (ICIP) | 978-1-6654-9620-9/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICIP46576.2022.9897859

ABSTRACT

Face recognition often suffers from severe degradation in accuracy when applied to images captured by fisheye cameras. One way to resolve the issue is to rectify the fisheye images before classification; however, it can only achieve local optimum. In this paper, we present an end-to-end model with global optimum. To tackle the challenges due to intra-class variance and diversity of fisheye transformations, we propose 1) a structural correction to guide the model learning, and 2) a spatial-transformer-networks embedded model to compensate for the non-linear distortion of fisheye lenses. We test the proposed model on the CelebA dataset and a real image dataset and achieve an average accuracy of 98.7% and 98.0%, respectively, which represent improvements of 4.05% and 5.72% over the state-of-the-art results.

***Index Terms*—**Fisheye image, fisheye distortion, face recognition, image rectification.

1. INTRODUCTION

Face recognition plays a critical role in various applications, including identity authentication, healthcare, etc. [1]. The advancement of neural network provides a promising direction for refining the classification of rectilinear face images. However, extending such methods to fisheye images introduces a performance degradation. The root cause is related to the non-linear transformation of fisheye lenses, which severely distorts facial features. One way to resolve the issue is to rectify the image before classification [2]. The main idea is shown in Fig. 1, which consists of two functional blocks: image restoration and identity classification.

However, the main drawback of the previous system is that it is not globally optimized. We observe that the two functional blocks described above are optimized separately by different objective functions. The restoration block aims to make the reconstructed image resemble a rectilinear image, and the classification block focuses on accurate identity classification. In machine learning, separate block optimization often yields sub-optimal solution. Therefore, we take a different approach that jointly optimize all functional blocks of a face recognition system.

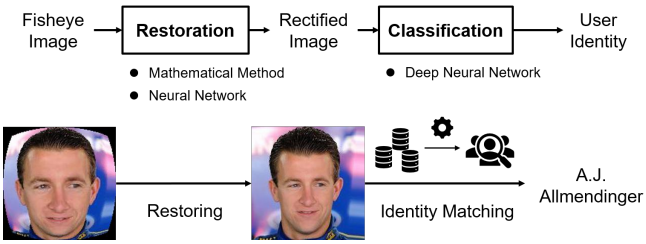

Fig. 1. System flow of the face recognition for fisheye images proposed by Li *et al.* [2].

In this paper, we propose an end-to-end training method for face recognition of fisheye images. We design two optimization methods based on two observations of severely distorted fisheye images. First, due to the diversity of fisheye distortions, the intra-class features of the images vary a lot and present a challenge to model training. Therefore, we add a structural correction component to the model by using an anchor to guide the model training and improve the clustering in feature of the same class. Second, the non-linear distortion introduced by the fisheye lens is radial symmetric, not translational symmetric, and it makes the model difficult to learn. To solve the problem, we adopt spatial transformer networks (STNs) in our face recognition model to make the hyper-features robust to fisheye distortions. We validate our model on the CelebA dataset [3] and a real image dataset. The results show significant improvements by 4.05% and 5.72% over previous methods.

2. RELATED WORK

Most previous research focused on the rectification of fisheye images, for which some rule-based methods rooted on mathematical theories have been developed. Ying *et al.* [4] proposed a Great Circle Fitting (GCF) algorithm to rectify fisheye distortion by fitting great circles in the fisheye image to straight lines. Huang *et al.* [5] determined the distortion function based on the radii of distorted circles in the fisheye images. However, these methods require prior measurements of the cameras, which may not be feasible in the real world.

Table 1. Comparison of different algorithms for face recognition on fisheye images.

Algorithms	w/o Prior Knowledge	Arbitrarily Clipped Image	End-to-end Optimization
Rule-based [4, 5]	×	×	×
FE-GAN [6]	◎	×	×
OD-FIR [2]	◎	◎	×
Proposed end-to-end network	◎	◎	◎

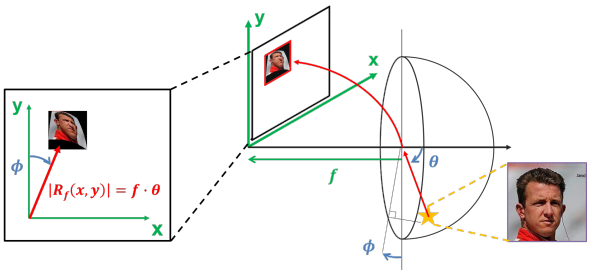

Fig. 2. Illustration of transformation between a rectilinear image and the clipped fisheye image.

The advancement of machine learning has helped improve the rectification quality. Yin *et al.* [7] rectified fisheye images by predicting the camera parameters and using semantic segmentation maps for image mapping. Li *et al.* [8] determined pixel-wise correspondences between the fisheye image and the distortion-free image and improved the rectification by re-sampling. Chao et al. [6] proposed an end-to-end training method for fisheye image rectification without any prior-knowledge. By combining Self-Supervised Learning and Generative Adversarial Networks (GANs), Fisheye GAN (FE-GAN) was able to deliver the state-of-the-art performance for fisheye image rectification.

On the other hand, Li *et al.* [2] focused on the development of a complete face recognition system and found that the cost function should address the final face recognition accuracy instead of the rectification quality. Although faces to be recognized came from any possible location of a fisheye image, they observed that not all face images require rectification. Therefore, they proposed an on-demand fisheye-image rectification (OD-FIR) framework, which only rectified the face images that may potentially affect the classification accuracy. The on-demand mechanism eases the reconstruction task and improves the overall accuracy and makes the state-of-the-art performance. We compare various face recognition systems in Table 1.

3. PROPOSED METHOD

We consider end-to-end training for recognition of face images extracted from fisheye images, similar to OD-FIR [2]. We propose two optimizations to overcome the distortion of the fisheye lens.

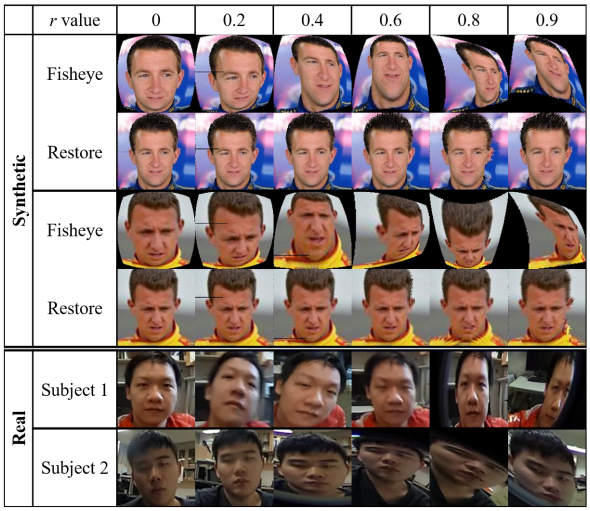

Fig. 3. Demonstration on clipped fisheye images, restored images, and real images with respect to different r values.

3.1. System model

We adopt the ArcFace [9] design as the base of our face recognition framework. The key of our design is that, in addition to applying the cross-entropy loss during training, the *arcface loss* is adopted to enhance the inter-class difference. By applying the arcface loss to the feature layer of the model, the intra-class features would converge into a cluster, and the angular distance between clusters can be expanded. The total loss can be formulated as follows:

$$\mathcal{L} = loss_{CE}(x, class) + loss_{AF}(f, class) \ , \qquad (1)$$

where $loss_{CE}(\cdot)$ denotes the cross-entropy loss, $loss_{AF}(\cdot)$ denotes the arcface loss, x denotes the output of the classifier, and f denotes the feature layer.

Then we train the recognition model by synthetic fisheye images. Fig. 2 demonstrates the generation of clipped fisheye images by projecting rectilinear images into the fisheye lens by an arbitrary incident angle (θ) and an arbitrary azimuth angle (ϕ). The image formulation function of the fisheye lens is

$$R = f \cdot \theta \ , \qquad (2)$$

where R denotes the radial distance of an image point to the center of the fisheye image, and f denotes the focal length of

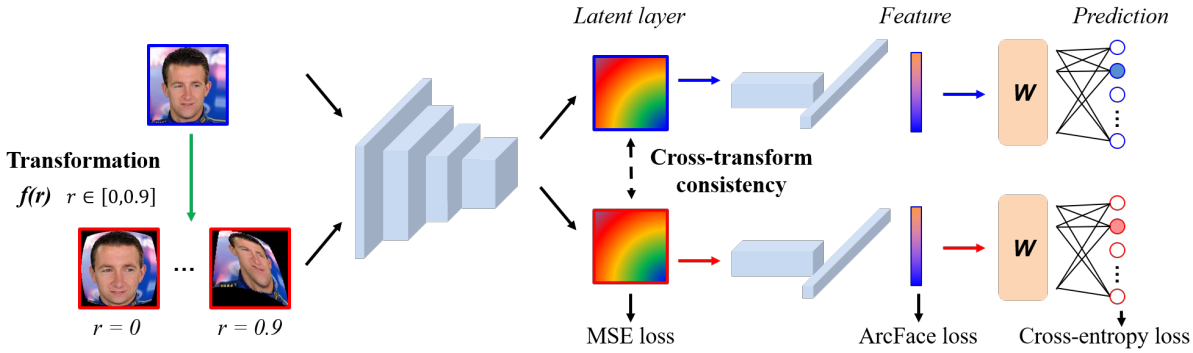

Fig. 4. Network structure of structural correction for fisheye image. During training, we input a clipped fisheye image and the corresponding rectilinear image into the model.

the fisheye lens. Notice that R is proportional to f, therefore, we normalize R into $r \in [0, 0.9]$. Fig. 3 shows the distorted results with respect to r. We can see that the distortion is mild when r is small, and becomes serious when r grows. Moreover, in addition to the randomly selected θ and ϕ, the distortion is non-translational symmetric. The diversity of distortion is the main challenge of the recognition task.

3.2. Structural correction on latent feature

We observe that the first challenge for face recognition of fisheye images lies in the intra-class variance issue. Fig. 3 illustrates a variety of distorted images. With the randomly selected θ and ϕ, the clipped fisheye images with identical origins may differ immensely from each other (images in a row of Fig. 3). The phenomenon leads to feature variation for the same identity, difficult for the model to classify.

To address the intra-class variance issue, we apply structural correction to the distorted features by providing a reference $anchor$. Fig. 4 shows our network architecture. In the training stage, aside from the clipped fisheye image, we provide the model with the corresponding rectilinear image as an anchor. By enforcing the latent feature of the clipped fisheye image to converge towards the anchor, the difficulty with intra-class features can be eased. We apply the *cross-transformation consistency* by adding the mean-square-error (MSE) loss during training defined as follows:

$$loss_{MSE}(f_f, f_r) = \|f_f - f_r\|_F \ . \tag{3}$$

f_f and f_r denote the latent features of the clipped fisheye image and the reference rectilinear image, respectively, and $\|\cdot\|_F$ denotes the Frobenius norm. Therefore, the training loss becomes

$$\begin{aligned}\mathcal{L} = \alpha \cdot loss_{CE}(x, class) + \beta \cdot loss_{AF}(f, class) \\ + \gamma \cdot loss_{MSE}(f_f, f_r) \ ,\end{aligned} \tag{4}$$

where α, β, and γ are adjustable parameters.

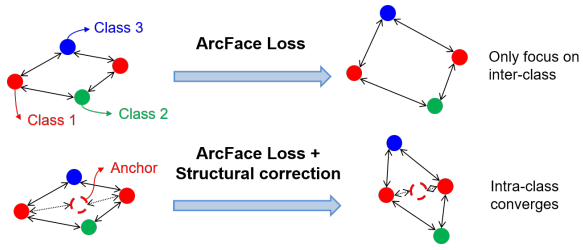

Fig. 5. Illustration for clustering on intra-class features using structural correction.

Fig. 5 illustrates the effect of the structural correction. With the ArcFace loss alone, the network separates the features of different classes. However, the features of the same class remain isolated and non-clustered, which hinders the classification. With the assistance of the structural correction, the distanced fisheye features have the anchor to converge toward. Therefore, the intra-class features can form a cluster easier and thus improve the accuracy.

3.3. Spatial-transformer-networks embedded model

The second challenge for classifying the clipped fisheye image is the diversity of the fisheye transformations. Fig. 3 shows the diversity of transformations and that the non-linear distortion is non-translational symmetric. Although the pooling layer in the neural network can mitigate the effect of distortion, the profound curvature transformation still cannot be solved by this process.

Inspired by the method proposed by Jaderberg *et al.* [10], we adopt Spatial Transformer Network (STN) to compensate for the fisheye distortion. We notice that the fisheye images are the projections of original rectilinear images. Hence, we add STNs into the intermediate layers of our model to learn the best projection for the features. The effect of fisheye distortion can be eased by the rectification on the latent feature maps.

Fig. 6 sketches our proposed network. We insert several

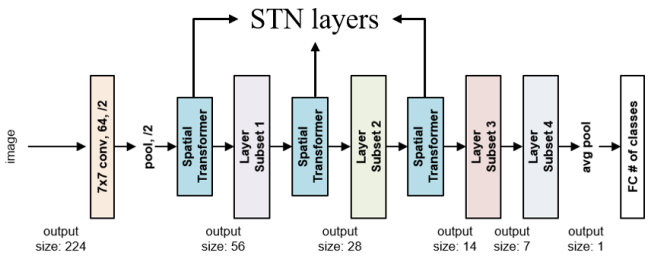

Fig. 6. Proposed spatial-transformer-networks (STNs) embedded model architecture.

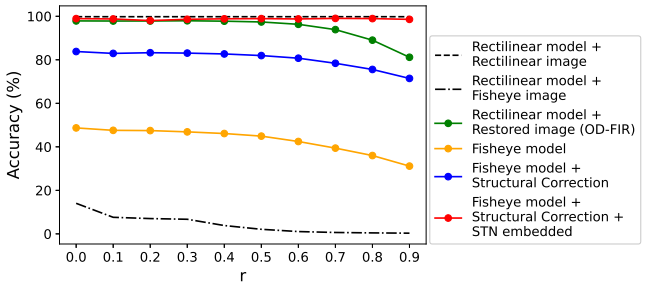

Fig. 7. Accuracy of different models and different input images.

STNs between sub-blocks of the backbone network to overcome the fisheye distortion. With the assistance of STNs, the distorted features can be gradually rectified.

4. EXPERIMENTS

4.1. Experimental setup

We first evaluate our method on the synthetic CelebA dataset [3]. We synthesize the clipped fisheye face image by transforming every image in the dataset by the fisheye-lens model shown in Section 3.1. We categorize the transformed images by different r values to observe the influence of r on the accuracy. In addition, we validate our method on a real image dataset illustrated in Fig. 3 to evaluate the proposed optimization methods. The dataset, which was acquired by a Ricoh Theta S camera, contains 50,000 images of 100 subjects.

We select Resnet34 [11] as our backbone network, and train it by four different settings: 1) The previous system (OD-FIR [2]) with a restoration block and a classification block in sequence. To avoid the algorithm-level bias, we assume all parameters of the camera and the projection angles available are perfect. 2) The end-to-end model for clipped fisheye images without any additional optimization. 3) The model with structural correction. 4) The model with structural correction and STNs embedded.

4.2. Experimental results

Fig. 7 shows the simulation results using the synthetic dataset. The upper bound and lower bound for the experiment can be acquired via validation on Model 1. The upper bound is the result of testing Model 1 with rectilinear face images, which hits an accuracy of 99.1%. The lower bound is the inference on the clipped fisheye images, for which the accuracy significantly drops to 14.0% and becomes more serious to 0.3% when r increases.

The reference bound is the performance of OD-FIR, which can be measured by validating Model 1 with restored fisheye images of different r values. The accuracy is shown in green in Fig. 7, where we can see that the performance has serious degradation when r is larger than 0.4. Clearly, this is the result of the fact that the fisheye lens extrudes the image and causes information loss. Since the restoration block in OD-FIR can be regarded as an over-sampling technique, such information loss cannot be compensated. Therefore, the accuracy inevitably drops when r becomes larger.

The accuracy of our end-to-end method under different conditions is shown by the orange, blue, and red lines in Fig. 7. Without any optimizations, Model 2 only uses the cross-entropy loss and the arcface loss to categorize the input image. With fisheye distortion, the intra-class difference is so significant that the network can not learn. Therefore, the average accuracy is only 43.1%. With the aid of structural correction, the cross-transform consistency can be held, hence the average accuracy is significantly improved to 80.4%. Furthermore, with the assistance from STNs, the average accuracy is further improved to 98.7%, which is only 0.4% lower than the upper bound and 4.05% higher than OD-FIR. Lastly, we evaluate our model on a real image dataset and get an average accuracy of 98.0%, or 5.72% improvement.

5. CONCLUSION

We investigate the feasibility of an end-to-end network for face recognition of fisheye images. To overcome the considerable intra-class variation caused by the fisheye lens, we apply structural correction to make the network learn the cross-transformation consistency. Furthermore, to compensate for the fisheye distortion, we adopt the STNs and insert them between the blocks of the backbone network. Our proposed optimizations achieve an accuracy of 98.7% for the synthetic CelebA dataset and 98.0% for a real image dataset.

6. ACKNOWLEDGMENT

This work was supported in part by grants from the Ministry of Science and Technology of Taiwan under Contracts 110-2221-E-002-108-MY3, 110-2221-E-002-184-MY3, and 110-2622-8-002-018 and in part by grant from Cathay United Bank under contract 109-3111-8-002-002.

7. REFERENCES

[1] Marius Drulea, Istvan Szakats, Andrei Vatavu, and Sergiu Nedevschi, "Omnidirectional stereo vision using fisheye lenses," in *2014 IEEE 10th International Conference on Intelligent Computer Communication and Processing (ICCP)*, 2014, pp. 251–258.

[2] Yi-Hsin Li, I-Chan Lo, and Homer H. Chen, "Deep face rectification for 360° dual-fisheye cameras," *IEEE Transactions on Image Processing*, vol. 30, pp. 264–276, 2021.

[3] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

[4] Xianghua Ying, Zhanyi Hu, and Hongbin Zha, "Fisheye lenses calibration using straight-line spherical perspective projection constraint," in *Asian Conference on Computer Vision*. Springer, 2006, pp. 61–70.

[5] Fuyu Huang, Yongzhong Wang, Xueju Shen, Chao Lin, and Yudan Chen, "Method for calibrating the fisheye distortion center," *Applied optics*, vol. 51, no. 34, pp. 8169–8176, 2012.

[6] Chun-Hao Chao, Pin-Lun Hsu, Hung-Yi Lee, and Yu-Chiang Frank Wang, "Self-supervised deep learning for fisheye image rectification," in *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2020, pp. 2248–2252.

[7] Xiaoqing Yin, Xinchao Wang, Jun Yu, Maojun Zhang, Pascal Fua, and Dacheng Tao, "Fisheyerecnet: A multi-context collaborative deep network for fisheye image rectification," in *Proceedings of the European conference on computer vision (ECCV)*, 2018, pp. 469–484.

[8] Xiaoyu Li, Bo Zhang, Pedro V Sander, and Jing Liao, "Blind geometric distortion correction on images through deep learning," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4855–4864.

[9] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2019, pp. 4690–4699.

[10] Max Jaderberg, Karen Simonyan, Andrew Zisserman, et al., "Spatial transformer networks," *Advances in neural information processing systems*, vol. 28, pp. 2017–2025, 2015.

[11] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2016, pp. 770–778.