ONE-SHOT DEEP NEURAL NETWORK FOR POSE AND ILLUMINATION NORMALIZATION FACE RECOGNITION

Zhongjun Wu, Weihong Deng

Beijing University of Posts and Telecommunications
No. 10, Xitu Cheng Road, Haidian District, Beijing, China, 100876
wuzhongjun1992@126.com, whdeng@bupt.edu.cn

ABSTRACT

Pose and illumination are considered as two main challenges that face recognition system encounters. In this paper, we consider face recognition problem across pose and illumination variations, given small amount of training samples and single sample per gallery (a.k.a., one shot classification). We combine the strength of 3D models in generating multi-views and various illumination samples and the ability of deep learning in learning non-linear transformation, which is very suitable for pose and illumination normalization, by using a multi-task deep neural network. By the pose and illumination augmentation strategy, we train a pose and illumination normalization neural network with much less training data compared to other methods. Experiments on MultiPIE database achieve competitive recognition results, demonstrating the effectiveness of proposed method.

Index Terms— face recognition, 3D model, one shot deep neural network, pose and illumination augmentation

1. INTRODUCTION

Face recognition techniques have made great process over past years for the huge potential in real world applications, such as access control and video surveillance. Pose and illumination are considered as two significant factors that dramatically affect the performance of face recognition system.

In face recognition, pose problem is essentially a misalignment problem caused by the rigid transformation of 3D face structure and in general, 3D based method are more precise and achieve higher accuracy than 2D methods. Illumination problem refers to the huge differences of face shading and shadow for varying direction and energy distribution of the ambient illumination together with the 3D structure of face. The complexity of face recognition increases when illumination variation is coupled with pose problem. Such variations sometimes are as large as or even larger than that caused by differences due to identity [1].

3D Morphable model (3DMM) [2] fits the parameters of 3D shape, texture, illumination and pose and use them as representation. But it is hard to implement it in practical system for high computational burden.

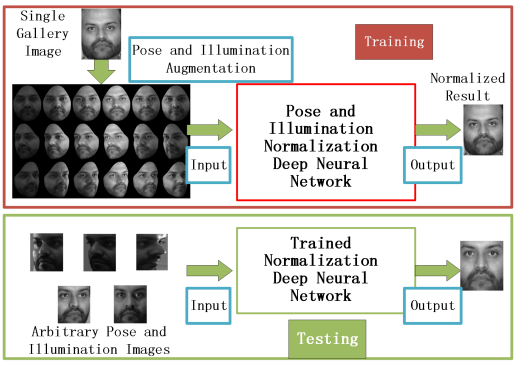


Fig. 1. Visual illustration of proposed method.

Recent years, deep learning techniques have swept a variety of computer vision tasks including face recognition with pose and illumination problem. The general idea is to convert a non-frontal, arbitrary lighting face to a canonical pose (usually frontal), neutral light one. Face identity-preserving (FIP) [4] utilizes face images with arbitrary pose and illumination variations as input and reconstructs corresponding face under frontal-view, neutral light to preserve identity information through a deep network. Multi-View Perceptron (MVP) [5] extends [4] which can infer a full spectrum of multi-view images and untangle the identity and view feature using random hidden neurons. Controlled Pose Feature (CPF) [6] is a recent work which can rotate the arbitrary pose, illumination, query face into several target pose faces as multiple classifiers in recognition, by using a multi-task deep neural network.

Deep learning techniques have strong ability in learning representation and non-linear transformation, which is very suitable for solving misalignment and self-occlusion problem in terms of pose, as well as for lighting transformation. However, tens of thousands of training samples, covering discrete poses and various lighting conditions are needed to learn the transformation function and performance is largely related to the amount of training image. The performance outside the sampled poses is also difficult to predict.

In this paper, we consider face recognition problem with pose and illumination variations, given small amount of training samples and single sample per gallery (a.k.a., one shot classification). The idea is to combine the strength of 3D models and the ability of deep learning in learning non-linear transformation.

In training stage, given a small 2D training set, for each gallery image, build corresponding 3D models virtually under different illumination conditions, based on a single 3D shape prior. Then render those 3D models under different views to synthesize images of different poses and illumination conditions. We use the above images as input and corresponding frontal, neutral light face from same identity as output of a neural network, aiming to learn a pose and illumination normalization neural network, which architecture is similar to CPF [6]. In test stage, a non-frontal, arbitrary lighting query face image is sent to the neural network and recognition is performed by distance comparison between the normalized output image and the gallery frontal images. Visual illustration is shown in Fig. 1.

The main contributions of this work can be summarized as follows. Considering the condition that small training samples can be acquired, we provide an idea of face multiple views and lightings augmentation using 3D model to train a deep neural network, with single sample per gallery. Experiments on MultiPIE database [7] achieve reasonable results and verify the effectiveness of our method, with much less training data.

The rest of paper is organized: Sec. 2 introduces the process of poses and illumination augmentation. In Sec. 3, the normalization network is described in detail. Sec. 4 is experiment part, followed by the conclusion part in Sec. 5.

2. FACE POSES AND ILLUMINATION AUGMENTATION

The idea can be summarized as: "From *one* to *many*". Concretely, given a set of frontal face images captured under various lighting conditions, with a query face, corresponding multi-view face images under same lighting conditions that training set contain are generated.

2.1. Single 3D Shape Prior

We assume that each face subject has similar 3D face shape that can be derived from a single 3D shape. That is, a single 3D shape can be transformed to a specific identity face shape. Our face re-rendering (a.k.a., relighting) method Adaptive Quotient Image (**AQI**) and the 3D face reconstruction method Generic Elastic Models (**GEMs**) share this shape prior and will be described below. Detail procedure can be seen in Fig. 2. Although this prior seems strong, experimental results show that coarse face shape is good enough to achieve reasonable results.

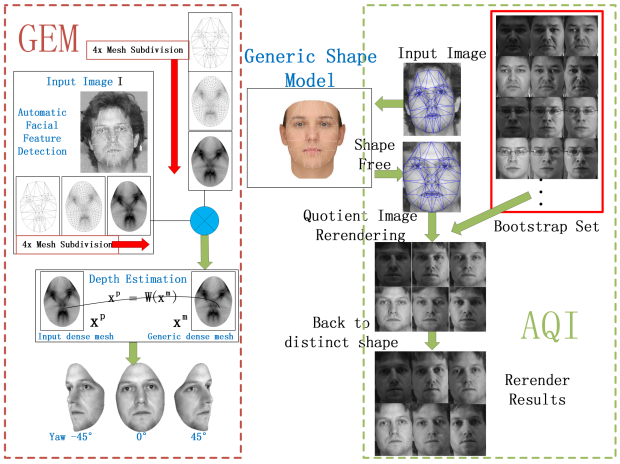


Fig. 2. Visual illustration of single shape prior. The face re-rendering method Adaptive Quotient Image (AQI) and 3D face reconstruction method 3D Generic Elastic Models (GEM) share this prior.

2.2. Adaptive Quotient Image

Quotient Image method [8] is a classical work of face re-rendering. Face, as a class of object, can be considered as Lambertian Surface with a reflection function: $\rho(x,y)n(x,y)^T s$, where $0 \leq \rho(x,y) \leq 1$ is the surface reflectance (gray-level) associated with point x,y in the image, $n(x,y)$ is the surface normal direction associated with point x,y in the image, and s is the (white) light source direction (point light source) and whose magnitude is the light source intensity.

In [8], the concept *Ideal Class of Object*, i.e., objects that have same shape but differ in surface albedo is defined. Under this assumption, the *Quotient Image* $Q_y(u,v)$ of face y against face a is defined: $Q_y(u,v) = \frac{\rho_y(u,v)}{\rho_a(u,v)}$, where u,v range over the image. Thus, Q_y depends only on the relative surface texture information and is independent of illumination.

A bootstrap set containing N (N is small) faces under M unknown independent illumination (totally $M \times N$ images) is adopted. Q_y of a input image $Y(u,v)$ can be calculated as $Q_y(u,v) = \frac{Y(u,v)}{\sum_{j=1}^{M} \bar{A}_j(u,v) x_j}$, where $\bar{A}_j(u,v)$ is the average of images under illumination j in the bootstrap set and x_j can be determined by the bootstrap set images and the input image $Y(u,v)$. The image space created by the input object y, under varying illumination, is spanned by the product of images Q_y and $\sum_j \bar{A}_j z_j$ for all choices of z_j.

The original rough face alignment process in [8] (the center of mass was aligned and scale was corrected manually) can not well satisfy the *Ideal Class of Object* definition. We think of a idea that if all images are transformed to a generic shape, the definition can be better satisfied. So we adapt the alignment process to warp all images into a generic shape by affine transformation and then perform Quotient Image method (see

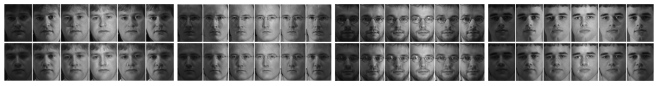


Fig. 3. Example face re-rendering results. First row: Faces captured under specific lightings; Second row: Corresponding re-rendered results by AQI.

the right part of Fig. 2). Example re-rendered results are demonstrated in Fig. 3. The result images of AQI are close to ground truth images.

2.3. 3D Generic Elastic Models (3D GEMs)

The topic of this part can be described as: "3D face reconstruction from a single 2D image". Shape From Shading [9], estimates depth information purely from images but the acquired model is not satisfactory enough and it requires face symmetry assumption. 3DMM is powerful enough but it suffers high computational complexity.

GEMs was introduced in [10, 11] as a low computational but efficient 3D modeling method. The underlying assumption is that face depth information does not dramatically change between individuals as long as the corresponding 2D face feature points are aligned. Let's recall the single shape prior. The idea of GEM is that by deforming a dense generic shape on the (x, y) plane, a 3D shape corresponding to input identity can be generated.

The procedure can be summarized: First, 77 sparse landmarks are detected on both input image and generic shape model. By Delaunay Triangulation, a sparse mesh is generated. Then by 4x mesh Loop Subdivision, dense correspondence is constructed. After depth assignment and texture mapping from input image, the 3D model is generated (see the left part in Fig. 2).

2.4. Conclusion of Pose and Illumination Augmentation

With Adaptive Quotient Image, we first virtually re-render the input image and construct corresponding 3D models from the re-rendered images by 3D Generic Elastic Models. These 3D models are rendered under different views to synthesize images under various views and illumination conditions. With the single 3D shape prior, we make pose and illumination augmentation with only one gallery sample, given a small set of training data.

3. POSE AND ILLUMINATION NORMALIZATION NETWORK

In this part, we describe the pose and illumination normalization neural network. We use a simplify architecture of the one proposed in CPF [6]. Unlike FIP which only have a normalization task, CPF introduced an auxiliary reconstruction task that reconstructs the original input image from the output of the normalization task, to improve the identity-preserving ability of the DNN. The idea is that the output of normalization task should be identity-preserving and contains sufficient information of input identity to reconstruct the input image.

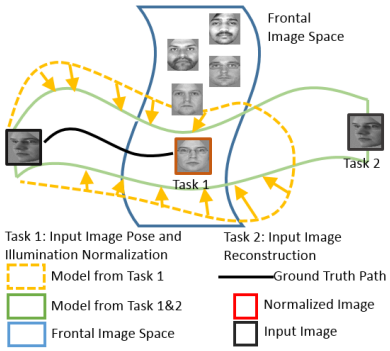


Fig. 4. Conceptual diagram of CPF Deep Neural Network.

The conceptual diagram is shown in Fig. 4. A DNN trained by a single normalization task would warp the input image along a path that deviates from the ground truth path, illustrated by the yellow region. With the auxiliary task, the yellow manifold would shrink to the green one, which is closer to the ground truth path and identity-preserving ability is improved.

3.1. Model Description

Since we just normalize the input image to frontal pose, we simplify the architecture of CPF [6] which contains several target pose in normalization task. The final network architecture is shown in Fig. 5.

The conventional CNN [12] uses shared weight filters. But it's inappropriate to apply it in this task for the reason that different regions of face need different kind of non-linear transformation. Therefore, locally connected layer without weight sharing is used. Early fully connected layer (the green layer in Fig. 5) can change features to contain pose and illumination information of input identity.

The whole parameters of DNN is described as Input(60×60)-L(7,32)-P(2,2)-FC(60×60)-L(5,32)-P(3,3)-FC(60×60+340)-FC(60×60)-L(5,32)-P(3,3)-FC(60×60). L, P, FC mean locally connected layer, max pooling layer, fully connected layer respectively. L(7,32), P(2,2), FC(60×60) mean that the layer applies 32 filters without weight sharing with the kernel size 7, the max pooling layer applies kernel size 2 with stride 2, the fully connected layer with 60×60 neurons. FC(60×60+340) is the normalized output layer (the red layer in Fig. 5) with normalized image (60×60) and the **Recon Code** (with length 340) which represents pose and illumination information of input image. The locally connected layer and fully connected layer use ReLU [12] as activation function. But the normalized output layer and reconstruction layer (the red and yellow layers in Fig. 5) do not apply activation function.

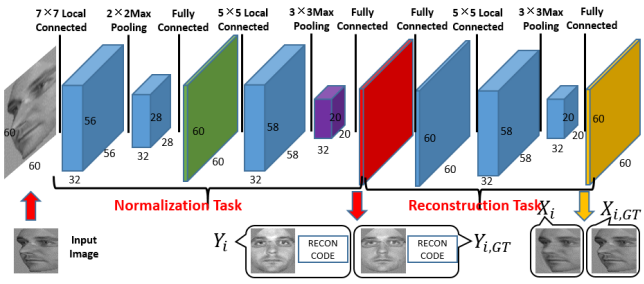


Fig. 5. Complete DNN architecture of our model.

3.1.1. *Recon Code*

Since the auxiliary DNN reconstructs the input image with the normalized results (frontal pose and neutral lighting), it needs both the pose and illumination information of input image to reconstruct it. So the output layer code, called **Recon Code** is set, to represents the i-th pose out of n poses with the t-th illumination condition out of m illumination variations of input image.

Recon Code contains two parts, pose code and illumination code. Pose code $Q_i \in \{0,1\}^l$ with total length l is defined as:

$$Q_i^j = \begin{cases} 1 & \text{if } (i-1)\times k < j \leq i \times k \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where Q_i^j is the j-th bit of code Q_i and $k = \lfloor l/n \rfloor$. (l,n) are set equal to $(140,7)$. Illumination code $S_t \in \{0,1\}^l$ with total length l is defined similarly as:

$$S_t^j = \begin{cases} 1 & \text{if } (t-1)\times k < j \leq t \times k \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

where S_t^j is the j-th bit of code S_t and $k = \lfloor l/m \rfloor$. (l,m) are set equal to $(200,20)$.

3.1.2. *Cost Function*

Square Euclidean distance loss is used for the two tasks. For the normalization task, cost function is defined as: $E_c = \sum_{i=1}^{N} \|Y_{i,GT} - Y_i\|_2^2$, where $Y_{i,GT}$ and Y_i are the ground-truth and output of normalization layer that contain normalized image and Recon Code. N indicates the training batch size. Similarly, the cost function of reconstruction task is: $E_r = \sum_{i=1}^{N} \|X_{i,GT} - X_i\|_2^2$, where $X_{i,GT}$ and X_i are the ground-truth and reconstruction result of input image.

Two tasks are considered as equal importance and the total cost function is: $E = \lambda_c E_c + \lambda_r Er$, where $\lambda_c = \lambda_r = 1$.

3.1.3. *Another Perspective towards this DNN*

This network can be viewed as a convolutional version of auto-encoder neural network [13]. Conventional auto-encoder network takes a vector as input and tries to reconstruct it by fully connected layers while this network applies locally connected layers, pooling layers, fully connected layers to reconstruct a 2D image convolutionally. With the designed constraint (normalization task), this network forces the non-linear transformation manifold to go through a certain point in the space (the frontal pose, neutral illumination image of input identity), which benefits the reconstruction task. It indicates that human knowledge prior can help the design of neural network by adding some constraints.

4. EXPERIMENTS AND RESULTS

We conduct face recognition experiments across pose and illumination on MultiPIE database [7], which contains 754204 images of 337 identities. Each identity has images captured under 15 poses and 20 illumination conditions in four sessions during different periods, supporting development of algorithms for face recognition across pose, illumination and expression.

The setting introduced in [3, 4, 6] is used, which adopts images in session *one* with totally 249 identities. Images from $-45°$ to $+45°$ (seven poses) under 20 illuminations (marked as ID 00-19) are used. Images of first 100 identities are for training, and the images of the remaining 149 identities for test. During test, one frontal image of each identity in the test set is selected in the gallery. The remaining images from $-45°$ to $+45°$ except $0°$ are selected as probes.

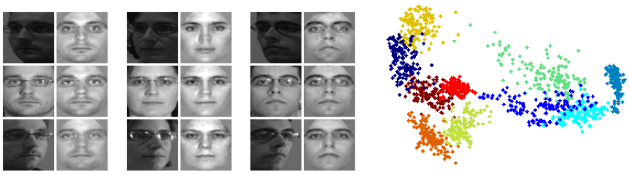
We extract features from the pooling layer in front of the normalized output layer (the purple layer in Fig. 5) to compare test image and gallery images by square euclidean distance and get recognition result.

4.1. Setting of proposed method

In our method, we empirically select frontal images of 12 identities from the first 100 identities (id 001, 002, 007, 008, 011, 012, 016, 019, 025, 026, 042, 047), under illuminations marked as ID 00-19 except 07, as the bootstrap set in AQI. Such small size bootstrap set is sufficient to achieve good re-rendering results. Different selection of bootstrap set hardly affects the re-rendering results.

Multi-Depth Generic Elastic Model (MD-GEM), introduced in [14], extends GEM by generating multiple 3D models for each input face based on the assumption that human face depth proportionally changes across people. We apply it to our pose and illumination augmentation framework to generate more samples that differ in face depth.

The normalization DNN is implemented by Caffe toolbox [15]. Our model is optimized using Stochastic Gradient Descent (SGD) with back propagation. The base learning rate, momentum, weight decay parameters are set to 0.001, 0.9, 0.04 respectively. The batch size is 90 and learning rate decreases exponentially through iterations. The input image is



(a) (b)

Fig. 6. (a) Example normalization results. (b) Normalized output feature space of 10 identities from normalization output layer (after PCA). Each identity has 140 (7 poses $\times$ 20 lighting) dots (one color per identity) and they converge to a certain region. Best viewed in color.

Table 1. Average Recognition Rate (Percent) on Different Poses. S1 and S2 refers to **Strategy 1** and **Strategy 2** respectively.

Methods	$-45°$	$-30°$	$-15°$	$+15°$	$+30°$	$+45°$	**Avg.**
Li [3]	63.5	69.3	79.7	75.6	71.6	54.6	69.3
RL [4]+LDA	67.1	74.6	86.1	83.3	75.3	61.8	74.7
CPF [6]	73.0	81.7	89.4	89.5	80.4	70.3	80.7
Ours GEM-S2	60.0	73.5	83.9	82.9	72.8	60.0	72.2
Ours MD-GEM-S1	40.4	54.2	70.8	72.6	55.1	42.4	55.9
Ours MD-GEM-S2	63.8	75.1	84.7	84.9	75.2	63.0	74.4

aligned to 60×60 pixels and subtracted and divided by the mean and variance of each image, for the robustness towards illumination variance.

Two training strategies have been tested.

Strategy 1: 12 identities as bootstrap set, use augmented images from the rest 88 identities of the first 100 identities to train the DNN, the non-frontal images from the remaining 149 identities for test.

Strategy 2: 12 identities as bootstrap set, use augmented images **directly** from the last 149 identities to train the DNN, the non-frontal images from the 149 identities for test.

4.2. Results

4.2.1. Feature Space

We extract features from the normalization layer of the test identities. Each identity has 140 probe images (7 poses $\times$ 20 illumination). We apply Principle Component Analysis (PCA) to all these features and randomly select features of 10 identities to show. As we can see from Fig. 6(b), the 140 normalized features from the same identity converge to a certain region, verifying the identity-preserving ability of our neural network.

4.2.2. Face Recognition Result

Table. 1 and Table. 2 report the results of our experiments. The recognition rate under a pose is the averaged result over all possible illuminations (marked as id 00-19, except 07).

Table 2. Average Recognition Rate (Percent) under Different Illuminations. S1 and S2 refers to **Strategy 1** and **Strategy 2** respectively.

Methods	00	01	02	03	04	05	06
Li [3]	51.5	49.2	55.7	62.7	79.5	88.3	97.5
RL [4]+LDA	72.8	75.8	75.8	75.7	75.7	75.7	75.7
CPF [6]	59.7	70.6	76.3	79.1	85.1	89.4	91.3
Ours GEM-S2	50.3	50.0	59.6	69.1	78.9	84.0	90.5
Ours MD-GEM-S1	37.6	35.0	43.4	51.2	58.5	68.6	76.1
Ours MD-GEM-S2	54.6	52.2	61.5	69.2	78.4	85.5	91.6
	08	09	10	11	12	13	14
Li [3]	97.7	91.0	79.0	64.8	54.3	47.7	67.3
RL [4]+LDA	75.7	75.7	75.7	75.7	75.7	75.7	73.4
CPF [6]	92.3	90.6	86.5	81.2	77.5	72.8	82.3
Ours GEM-S2	91.3	88.0	80.0	71.1	64.1	51.9	74.2
Ours MD-GEM-S1	78.9	73.6	64.5	52.9	42.5	37.6	57.0
Ours MD-GEM-S2	91.5	88.7	81.5	73.3	67.0	58.6	78.1
	15	16	17	18	19	**Avg.**	
Li [3]	67.7	75.5	69.5	67.3	50.8	69.3	
RL [4]+LDA	73.4	73.4	73.4	72.9	72.9	74.7	
CPF [6]	84.2	86.5	85.9	82.9	59.2	80.7	
Ours GEM-S2	78.5	82.1	79.8	77.2	50.9	72.2	
Ours MD-GEM-S1	60.7	67.1	61.7	57.7	37.6	55.9	
Ours MD-GEM-S2	79.8	84.6	83.1	79.8	55.0	74.4	

Table 3. Comparison of Training Images

Methods	Training Images
Li [3] RL [4]+LDA CPF [6]	100 Identities, 7 Poses, 20 Illuminations, Totally 14000 Images
Ours	12 Identities, Frontal Pose, 20 Illuminations, Totally 240 Images

The recognition rate under one illumination condition is defined similarly. We compare our results to three existing state-of-the-art face recognition across pose and illumination methods, Li *et al*. [3], FIP [4], and CPF [6].

The overall performance of MD-GEM-S2, 74.4% is as good as the one 74.7% in [4] and lower than the one 80.7% in CPF. It is reasonable because CPF uses several target pose images as multiple classifiers and achieves more reliable results while we uses one normalized output as single classifier.

As we can see, overall recognition rate of Strategy 1 is relatively low, 55.9%, compared to Strategy 2, 74.4%. It is because there are still gaps between our synthesized images and real images, in terms of illumination and side view. Since 3D models of all identities use identical face depth, synthesized non-frontal images share similarity in face depth. In Strategy 1, our network learns not accurate enough non-linear transformation through these shape-similar images. When we input real images from other identities, the not accurate enough transformation leads to degradation of identity-preserving ability. In Strategy 2, we directly train our network from the gallery identities, that is to say, the network "overfits" on the gallery identities and boost the performance, but just adopt one frontal, neural light image per gallery.

It is mentionable that our method doesn't use any non-frontal images from database to learn the pose information.

We just need a few frontal images under different illumination conditions as bootstrap set to learn illumination information (see Table. 3 for comparison). Pose information is learnt by 3D GEM based on a shape prior. We achieve competitive results with much less training samples (20 illumination $\times$ 12 identities $= 240$), compared to other methods which need tens of thousands of training samples to learn the pose and illumination transformation. Example normalization results are shown in Fig. 6(a).

Also, by using MD-GEM-S2, we boost the performance from GEM-S2, 72.2% to 74.4%, verifying the superiority of MD-GEM over GEM on 3D face reconstruction, which can be considered as a kind of face depth augmentation.

4.3. Limitations

In Table. 2, the performances of illumination id 00, 01, 13, 19 are relatively low. These conditions are captured under dark ambient light or strong specular light. It indicates that the limitation of Quotient Image with Lambertian surface assumption that can not model cast shadow or extreme dark ambient light very well. The re-rendering results are not satisfactory enough to some degree.

3D GEM is not accurate enough which leads to normalization result degradation when the actual shape of input identity is largely different from the generic shape model. Efficient, accurate 3D face reconstruction from several images or single image remains a challenging problem.

5. CONCLUSIONS

Pose and illumination are two main factors that affect performance of face recognition system. In this paper, we consider face recognition across pose and illumination problem, given a small set of training samples and single sample per gallery. We combine the strength of 3D models in generating multi-views and various illumination samples and the ability of deep neural network in learning non-linear transformation, which is very suitable for pose and illumination normalization. This augmentation idea can be applied when we train a deep neutral network but training samples are hard to acquire. Experiments on MultiPIE database achieve competitive results with much less training data, compared to other methods, verifying effectiveness of proposed method. In the future, we will consider more sophisticated work on combining 3D model and deep learning.

Acknowledgments

This work was partially sponsored by supported by the NSFC (National Natural Science Foundation of China) under Grant No. 61375031, No. 61573068, No. 61471048, and No.61273217, the Fundamental Research Funds for the Central Universities under Grant No. 2014ZD03-01, This work was also supported by the Beijing Higher Education Young Elite Teacher Program, Beijing Nova Program, CCF-Tencent Open Research Fund, and the Program for New Century Excellent Talents in University.

6. REFERENCES

[1] Xuan Zou, Josef Kittler, and Kieron Messer, "Illumination invariant face recognition: A survey," in *Biometrics: Theory, Applications, and Systems, 2007. BTAS 2007. First IEEE International Conference on*. IEEE, 2007, pp. 1–8.

[2] Volker Blanz and Thomas Vetter, "Face recognition based on fitting a 3d morphable model," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 25, no. 9, pp. 1063–1074, 2003.

[3] Annan Li, Shiguang Shan, and Wen Gao, "Coupled bias–variance tradeoff for cross-pose face recognition," *Image Processing, IEEE Transactions on*, vol. 21, no. 1, pp. 305–315, 2012.

[4] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Deep learning identity-preserving face space," in *Computer Vision (ICCV), 2013 IEEE International Conference on*. IEEE, 2013, pp. 113–120.

[5] Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang, "Multi-view perceptron: a deep model for learning face identity and view representations," in *Advances in Neural Information Processing Systems*, 2014, pp. 217–225.

[6] Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Du-Sik Park, and Junmo Kim, "Rotating your face using multi-task deep neural network," in *IEEE Conference on Computer Vision and Pattern Recognition, CVPR*, 2015, pp. 676–684.

[7] Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker, "Multi-pie," *Image and Vision Computing*, vol. 28, no. 5, pp. 807–813, 2010.

[8] Amnon Shashua and Tammy Riklin-Raviv, "The quotient image: Class-based re-rendering and recognition with varying illuminations," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 23, no. 2, pp. 129–139, 2001.

[9] Ruo Zhang, Ping-Sing Tsai, James Edwin Cryer, and Mubarak Shah, "Shape-from-shading: a survey," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 21, no. 8, pp. 690–706, 1999.

[10] Jingu Heo, *3D Generic Elastic Models for 2D Pose Synthesis and Face Recognition*, Ph.D. thesis, Citeseer, 2009.

[11] Utsav Prabhu, Jingu Heo, and Marios Savvides, "Unconstrained pose-invariant face recognition using 3d generic elastic models," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 33, no. 10, pp. 1952–1961, 2011.

[12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, 2012, pp. 1097–1105.

[13] Yoshua Bengio, "Learning deep architectures for ai," *Foundations and trends® in Machine Learning*, vol. 2, no. 1, pp. 1–127, 2009.

[14] Zhongjun Wu, Jiayu Li, Jiani Hu, and Weihong Deng, "Pose-invariant face recognition using 3d multi-depth generic elastic models," in *Automatic Face and Gesture Recognition (FG), 2015 11th IEEE International Conference and Workshops on*. IEEE, 2015, pp. 1–6.

[15] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the ACM International Conference on Multimedia*. ACM, 2014, pp. 675–678.