Deep Learning in Face Recognition Across Variations in Pose and Illumination

Xiaoyue Jiang, Yaping Hou, Dong Zhang, and Xiaoyi Feng

Abstract Even though face recognition in frontal view and normal lighting conditions works very well, the performance drops sharply in extreme conditions. Recently there is plenty of work dealing with pose and illumination problems, respectively. However both the lighting and pose variations always happen simultaneously in general conditions, and consequently we propose an end-to-end face recognition algorithm to deal with two variations at the same time based on convolutional neural networks. In order to achieve better performance, we extract discriminative nonlinear features that are invariant to pose and illumination. We propose to use the 1×1 convolutional kernels to extract the local features. Furthermore a parallel multi-stream convolutional neural network is developed to extract multi-hierarchy features which are more efficient than single-scale features. In the experiments we obtain the average face recognition rate of 96.9% on MultiPIE dataset. Even for profile position, the average recognition rate is also around 98.5% in different lighting conditions, which improves the state-of-the-art face recognition across poses and illumination by 7.5%.

1 Introduction

Face recognition has been one of the most active research topics in computer vision for more than three decades. With years of efforts, promising results have been achieved for automatic face recognition in both controlled [60] and uncontrolled environments [11, 17]. A number of algorithms have been developed for face recognition with wide variations in view and illumination, respectively. Yet few attempts have been made to tackle face recognition problems with the variations of pose and illumination [71]. In fact, face recognition is significantly affected by both pose and illumination which are often encountered in real-world images.

X. Jiang (✉) · Y. Hou · D. Zhang · X. Feng
School of Electronics and Information, Northwestern Polytechnical University, Xi'an, China
e-mail: xjiang@nwpu.edu.cn

X. Jiang et al. (eds.), *Deep Learning in Object Detection and Recognition*,
https://doi.org/10.1007/978-981-10-5152-4_3

Recognizing faces reliably across pose and illumination has been proved to be a much more difficult problem.

Pose can induce dramatic variations in face images. Essentially, this is caused by the complex 3D geometrical structure of the human head. The rigid rotation of head results in self-occlusion which means that some facial appearance will be invisible. At the same time, the shape and position of the visible part of facial images also vary nonlinearly from pose to pose. Consequently, the appearance diversity caused by pose is usually greater than that caused by identity. Thus general face recognition algorithms always fail when dealing with the images of different poses.

Illumination also can cause dramatic variations for face images. Assuming Lambertian reflectance, the intensity value $I(x, y)$ of every pixel in an image is the product of the incident lighting $L(x, y)$ and the reflectance $R(x, y)$ at that point as $I(x, y) = R(x, y) \times L(x, y)$. Thus, the captured images vary with the incident lighting. In order to achieve face recognition across illumination, there are two kinds of strategies. One is to extract illumination-invariant features from images, such as LBP [1] and HOG [65] et al.; the other is to model the distribution of illumination [24, 30].

In applications, both the pose and illumination variations exist. Thus a robust face recognition system should be able to deal with the two variations at the same time. Recently, the deep learning methods [27, 74] showed its great ability to model nonlinear distributions of data. It achieved the state-of-the-art performance in many fields of pattern recognition, such as object classification [58] and object detection [46]. Its great capacity is mainly due to the learning procedure which can find the hierarchical features from dataset. These features from each layer of the networks contain different levels of structure from a local gradient to its global shape. As a result, these learned features are more informative than traditional human-engineered features.

Even though different poses can induce the different appearances of the face, there exist some correlations between images of the same identity in different poses. Similarly, images of the same identity in different illuminations also correlate to each. Thus, through a proper learning method, the pose and illumination-invariant features can be obtained. Inspired by the excellent feature learning ability of deep convolutional neural networks, it is employed to develop an end-to-end face recognition method across pose and illumination in this work.

The remainder of this chapter is organized as follows: Sect. 2 briefly reviews the recent algorithms that deal with the pose problem. Section 3 introduces the algorithms that deal with the illumination variations. The proposed deep learning algorithm that can verify faces under pose and illumination variation is described in Sect. 4. The experimental results of the proposed algorithm are presented in Sect. 5. Finally, Sect. 6 concludes the chapter.

2 Pose-Invariant Face Recognition

Pose always causes substantial variations in the appearance of images due to the reason that images are the projection of 3D objects to 2D planar. Therefore when the pose of an object changes slightly, the appearance of the image will change dramatically. Consequently pose always brings difficulties for face recognition systems where the pose variation is unavoidable in uncontrolled environment. As a result, pose becomes one of the essential challenges for face recognition. Nowadays, researches also pay notable attention to deal with the pose variation problems. We can classify all the algorithms about pose variation into two categories: invariant representation-based algorithms and the model-based algorithms. For first category, invariant features or subspaces are constructed where the pose variation is removed, while the second type of algorithms tries to build up a generative model to predict the appearance of the object in different views.

2.1 Invariant Representation

In the classical frontal face recognition algorithms, face is always considered as a whole component. Therefore a lot of holistic approaches achieved quite good results. Principal component analysis (PCA) [62] is applied to find the eigenspace of face images; therefore face images can be represented by the projection values on those eigenvectors. Through the analysis, the dimension of face images has been reduced significantly, and the recognition is performed due to the distance in eigenspace. In fact, the assumption for these holistic approaches is that face position is fixed. Therefore, these algorithms try to find the relationship between corresponding pixel pairs among images. For the same person, corresponding pixels should have similar features, and the overall distance between images of the same subject is relatively smaller than that of different subjects. However, when the pose of the subject changes, the position of face components varies as well; consequently the correlation between corresponding pixel pairs is broken. Thus the holistic approach is no longer suitable for pose problem, but local components or features show their effectiveness in handling with the pose problem.

2.1.1 Engineering Designed Features

Landmarks (such as eyes, nose, and mouth) are the key points on face, which represent the key components in a face, as shown in Fig. 1. If the transform between the corresponding landmark points can be defined, then the same transform formula can be applied to convert two images. In order to find the same landmarks in images of different views, some robust feature extractors are used to describe the landmark points on faces.

Fig. 1 The landmarks on facial images. Five landmarks are labeled in each image. The transform formula can be calculated from the relationship between the corresponding landmarks

Scale-invariant feature transform (SIFT) feature [37] is widely used in computer vision tasks for the extraction of robust features. It has also been used for face recognition. In order to find the connections of images for different poses, Biswas et al. [8] extracted SIFT features for landmark points, which can provide rotation and scale invariant features. Then tensor analysis was applied to learn the transform matrix between the landmarks of different poses. With this transformation, images taken in different views can be converted to the frontal view to compare with the frontal probing images for verification. Also, local binary pattern (LBP) is a descriptor that finds great success in texture analysis. It computes the distribution of local region variance and encodes the distribution into numbers, which are very efficient for further pattern analysis. LBP is also applied to extracted features for local regions around landmark points. Then all the local region features are connected into a new feature, which becomes pose-invariant [10].

In fact, accurate landmark detection itself is also a challenging problem. Besides extracting robust features around landmark points, researchers also tried to define some key points to find correspondence between images of different views. Dreuw et al. [6] propose to use speeded-up robust feature (SURF) to extract features in dense grid, and then RANSAC method is used to find the matching points between images of varied view for face recognition. Liao et al. [36] propose a partial face recognition method without alignment. First, they apply SIFT-like descriptors to extract key points from facial images. And then for those key points, sparse representation method is used to build up a complicated dictionary for all the possible local facial regions around key points based on training images. The key idea behind this method is also to extract robust features for key facial components but omit their locations.

Region-based pose-invariant feature extraction methods are also explored recently. Without the locations of key points, local regions are considered as the basic unit to contain key facial components. Ahonen et al. [2] propose to divide images into subregions, as shown in Fig. 2. Then they extract LBP features for each subregion of a face. Within a subregion, the location information is omitted; only the texture feature is extracted by the LBP descriptor. Thus the extracted feature is robust to pose variations as long as the key components are still located in the same subregion. It is reported that the proposed face recognition algorithm can keep good performance when the rotation angle is within 15°. For large pose variations, the content of each subregion changes greatly; consequently the correlation between subregions is broken.

Fig. 2 Facial images are divided into local regions. Features are extracted based on each patch, thus the patch-based feature is invariant to pose

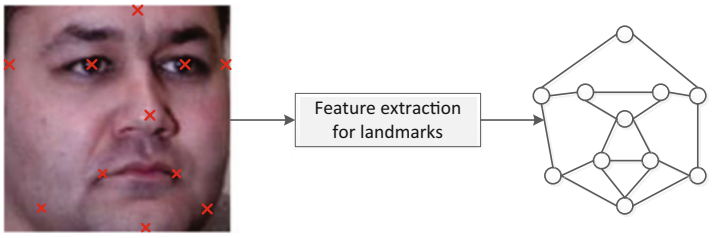

Fig. 3 Elastic graph for face recognition. The nodes of the graph are features extracted from local landmarks, and the edges of the graph represent the distance between neighboring nodes

Li et al. [33] propose a local region-based elastic matching method for face recognition across poses. Local descriptors, such as LBP or SIFT, are used to extract features for densely sampled subregions of images. The Gaussian mixture model is trained to extract the spatial-appearance distributions from the position of each local patch and its local feature. Each Gaussian model describes the relationship between corresponding patches of matched images. Then the verification is performed by a trained SVM that can discriminate the difference of Gaussian model between the matched and non-matched face images. In fact, the idea of elastic matching for face recognition of different poses is proposed by Wiskott et al. [67]. For each landmark of faces, a set of Gabor filters are applied to extract features, and then a graph with N nodes and E edges is constructed, where the nodes represent landmarks, and the edges are the feature distance between neighbored nodes, as shown in Fig. 3. Then the recognition is performed by comparing the graph of a probing face images to all the graphs of gallery images. The elastic bunch graph matching method can handle the rotation within 20°. For the elastic matching-based methods, a graph connects local components where the position of each component is also described by the graph. Thus the face can still be verified even though some local components are occluded.

Based on the cost of pixel-wised stereo matching, Castillo et al. [9] propose to do face recognition across poses. First they find three to four landmarks from face images to calculate the epipolar geometry parameters for gallery images. Then a stereo matching method is applied to find corresponding pixels between images.

Finally the cost of matching is used to identify the face images. Actually, there is an assumption that the corresponding pixels or components exist in two images. Thus when the pose changes greatly, the number of corresponding pixels between images decreases. Consequently, the performance of the algorithm drops significantly.

For the methods based on engineering-designed features, they try to find the corresponding local components between images. However, when poses change greatly, these manually designed features will always fail. The appearance of local components always varies greatly due to occlusion or out-of-plane rotation. Then the nonlinear correspondence should be found to describe the relationships.

2.1.2 Learning-Based Features

In order to find the nonlinear correspondence between images of different poses, some machine learning-based methods have been applied widely. Subspace learning methods are introduced to learn a new subspace that is invariant to pose variations. Metric learning methods are proposed to construct new distance measure methods, which are independent to the pose changes. Most recently, the deep neural network is also introduced to learn high-order nonlinear descriptors for images from different poses.

Linear Subspace Learning

In early years, principal component analysis (PCA) provides an important tool for extracting common features from dataset. The eigenvector that has the largest eigenvalue represents the direction of the biggest variance of the data, while the eigenvectors can be seen as the features shared among dataset. PCA-based methods achieve good performance in face recognition but are very sensitive to the misalignment of images. When face images are taken from different views, PCA encodes both identity and viewing conditions, which makes the performance of recognition degraded. Pentland et al. [42] propose to setup eigenspace for each component of the face, which only encodes the identity information, and the pose variance is alleviated by the selection of face components from images. When doing recognition, the reconstruction coefficients from each modular eigenspace are connected as a whole feature of the face.

Prince et al. [44] propose a statistical method to describe the distribution of face images regardless of pose. In the observed space, images from different views are located in different positions, where the difference caused by posture is much bigger than that caused by different identities. Thus it brings great difficulties for recognition. However, with the assumption that all faces of a single person in different poses can be described by a vector in the identity space, a linear transform mapping from the observation space to the identity space is proposed. Figure 4 shows the relationship between the two spaces. In the identity space, the pose

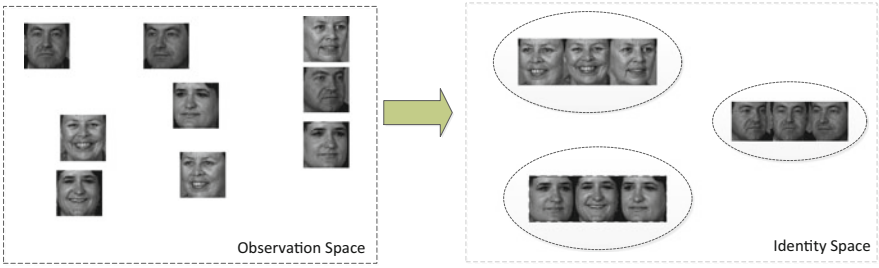

Fig. 4 Identity space only keeps the variance between different subjects but minimize the pose variance for a same subject

variance is diminished. Images from any pose can be represented as the linear combination of vectors in identity space h_i, and Gaussian noises ϵ_{ij} as follows:

$$X_{ij} = W_j h_i + \mu_i + \epsilon_{ij} \tag{1}$$

where W_j are the projection from identity space to the observation space and μ_i is the offset.

Therefore some generative models are introduced; the discriminative models are also used for dealing the pose problem actually. For discriminative models, they try to distinguish the difference between subjects regardless of pose variations. That is to maximize the margin between subjects or to find an optimal superplane to separate subjects.

Li et al. [31] propose to use canonical correlation analysis (CCA) to find a common space for images from different view, where the correlation between the same subjects is maximized, but not the traditional Euclidean distance. Correlation measures the difference of data tendency but not absolute distance. Thus it can allow some variance for data and becomes more robust to slight changes. The transform can be written as

$$\arg\max_{\omega_1, \omega_2} corr[\omega_1^T X_1,\ \omega_2^T X_1] \tag{2}$$

where $\|\omega_1\| = 1$, $\|\omega_2\| = 1$. X_i are images from different view and ω_i is the optimal transformation that can be solved by Lagrange multiplier method. In order to project images from all different poses to the same subspace but not only two views as CCA methods, Rupnik et al. [47] propose Multiview CCA(MCCA), as

$$\arg\max_{\omega_i, \ldots, \omega_k} \sum_{i \neq j} corr[\omega_i^T X_i,\ \omega_j^T X_j] \tag{3}$$

where $\|\omega_i\| = 1, i = 1, \ldots, k$. The set of transform ω_i can transform images from different views to the same subspace and meanwhile keep the maximal correlation

for the images of the same person. Based on MCCA, Sharma et al. [53] improve the algorithm further. They first find the optimal mapping functions for images from different views with the MCCA method, and then linear discriminative analysis (LDA) is applied to classify each subject in the new subspace.

Sharma et al. [52] propose a unified framework for multiview analysis called generalized multiview LDA (GMLDA), in addition. Within this framework, the discriminative analysis for each subject and the self-correlation of the images from the same subject are combined as

$$\arg\max_{\omega_i} \sum_i \mu_i \omega_i^T S_{bi} \omega_i + \sum_{i \neq j} \lambda_{ij} \omega_i^T Z_i Z_j \omega_i \tag{4}$$

where $\sum_i \gamma_i \omega_i^T S_{wi} \omega_i = 1$, μ_i, λ_{ij} and γ_i are parameters for linear combination. S_{bi} and S_{wi} are the between-class and within-class scatter matrix for the ith pose. The first term performs the LDA analysis for different poses, which enhances the distinguishability. On the other hand, the second term focuses on the correlation for all the images from the same subject. That is, the projection from different poses should be close to each other in the latent space. Z_i are the matrices whose column contains images of the same subject.

For CCA methods, it requires each subject to have exactly the same training data for each poses, while GMLDA only requires pairwise training data from different poses. In order to alleviate the requirement for training data, Kan et al. [28] proposed multiview discriminant analysis (MvDA). They apply the idea of LDA to analyze pose problem, where the intrapose scatter is minimized and interpose scatter is maximized. Then all the images from the same pose will be cluttered together in the new subspace.

Nonlinear Subspace Learning

The pose variations for images are due to the projection of 3D structure of the object to the 2D planar. As a result, poses actually bring nonlinear transform for the appearance of images. Consequently the nonlinear models should be more suitable than linear models for the description of pose variations. Kernel-based methods are the direct extension of linear methods, where the kernel can transform a linear subspace to a nonlinear subspace. In a nonlinear subspace, the nonlinear distributed data can be separated by a linear surface. Consequently, the classification can be achieved by linear methods in a higher-dimensional subspace.

There are quite a few adaptations to the linear methods, such as the kernel-PCA [49] is the extension of PCA by kernel method. Yang et al. [69] proposed a kernel Fisher discriminant framework by full usage of the KPCA and LDA. Experiments show the improvement in face recognition tasks. Recently, Sharma et al. [55] proposed a generalized multiview analysis (GMA) method which projects

a pair of images from different views into a common space by using the kernel tricks. The proposed method shows its effectiveness in pose-invariant face recognition.

Metric Learning

Besides using a suitable subspace to represent all the images from different poses, the distance metric also can be adapted to deal with pose variations. Schroff et al. [51] propose to compare the similarity of probe images with a big set of gallery images, and the similarity list is used as a feature to determine the identify of probe images. It is based on the assumption that images from the same person should have more common look-alike samples than that from different people, even if the images are taken in different conditions. Liao et al. [35] propose to compare the low-frequency information of the probe image with all the gallery images, and then a pooling method is applied to make it pose-invariant. Furthermore, Kafai et al. [25] construct reference face graph (RFG) to represent the relationship between different subjects, where each node in the graph contains all the images taken in different conditions of that subject, as shown in Fig. 5. The importance of each node is readjusted by its node centrality including degree, betweenness, and closeness for weighted graphs. Finally, the probe image is represented by the vectors that are composed of the similarity measure to each node in the graph where the hashing code is calculated for each oversampled region of images.

Fig. 5 In the reference face graph, each node is composed of faces from different views. Then all the faces are used as the basis to represent the probe face where the reference face descriptors are calculated

Deep Learning

Recently, deep learning-based methods show great success in the field of signal processing. It achieves the state-of-the-art performance in many applications such as face recognition, object classification, and so on. The great ability behind the neural networks is the nonlinear modeling capability actually. The nonlinearity is due to the nonlinear activation neurons in the networks. In addition, the multiple layers of the neural network make the order of the nonlinear model much higher than traditional models, which can represent the data more accurately.

Andrew et al. [3] proposed to use two parallel neural networks for the feature extraction of images from different poses, and then the output items from the networks are maximally correlated, where the correlation value is used to optimize the parameters of the neural networks. In fact, deep networks are performed as nonlinear mapping functions for input images. As a result, the deep canonical correlation analysis (DCCA) method shows better performance than Kernel CCA and CCA in the experiments, which is due to the robust features extracted by neural networks. The structure of the DCCA is shown in Fig. 6, where a three-layer fully connected neural network is applied. With the improvement of the neural network, more sophisticated features can be extracted.

Zhu et al. [73] propose a deep network to find the identity preserving features from images in different views, as shown in Fig. 7. There are three convolutional layers in the deep network. The input of the network can be images from any

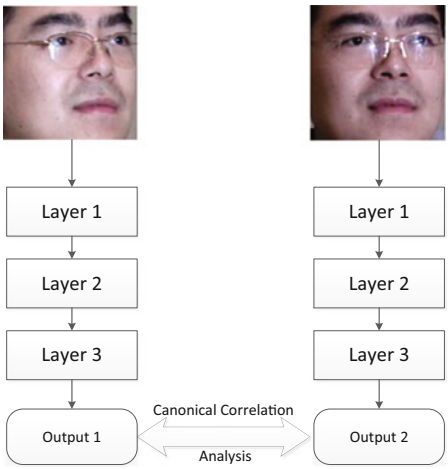

Fig. 6 Deep canonical correlation analysis for images from different poses. Two parallel neural networks are used to extract features for images from different poses; the output features are constrained to be maximally correlated to each other

Fig. 7 Identity preserve network. The neural network is applied to reconstruct the frontal view of input images from different poses

poses, while the output of the network is the reconstruction of the frontal view for the input subject. The network is composed of two basic units, which are the feature extractor and the reconstructor for the frontal view. In training, the difference between the reconstructed images and the ground truth is calculated for back propagation. Through the supervised learning procedure, the network can recover any input image to its corresponding frontal view. The extracted features also show great capability for identity discrimination. The face recognition is performed by comparing the recovered frontal face with the gallery frontal images. For this network, the last reconstruction layer is achieved by fully connection layer, which has millions of parameters needed to be trained. Consequently, it requires a huge number of images for training.

Based on the encoder framework, Zhang et al. [70] propose an encoder network for images of different poses. There is only one hidden layer for this network. The input of this network is images from different poses, and the output is the frontal image of the same subject. The encoder tries to find common features for images from different views. Furthermore, they propose to use random images to represent identity of each subject and train the encoder to find discriminative features for each subject. In order to keep the convergence of training, the sparse constraint of parameters is added to the loss function. Figure 8 shows the structure of the network. Compared with Zhu's work [73], it only has one hidden layer, which reduces the number of parameter but also reduces the capability of model. It also proves that increasing the layers of neural network can enhance the order of nonlinearity of the model, which can increase the discriminability of the model. Kan et al. [26] also noticed that only one hidden layer is not enough to model the nonlinear transform from any pose to frontal view. Thus they propose to use a cascade of autoencoders to transform the pose gradually from non-frontal view to frontal view, as shown

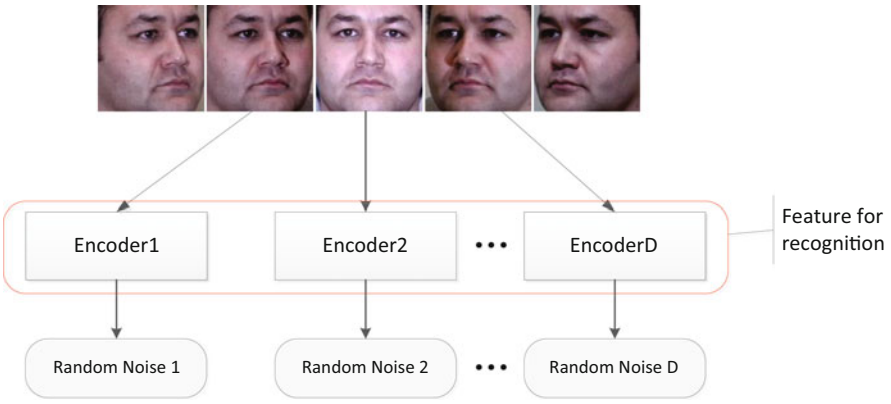

Fig. 8 Random face learning network. The random images are used as the output of one-layer encoder to learn the features for images from different views. Then the learned features are used for recognition

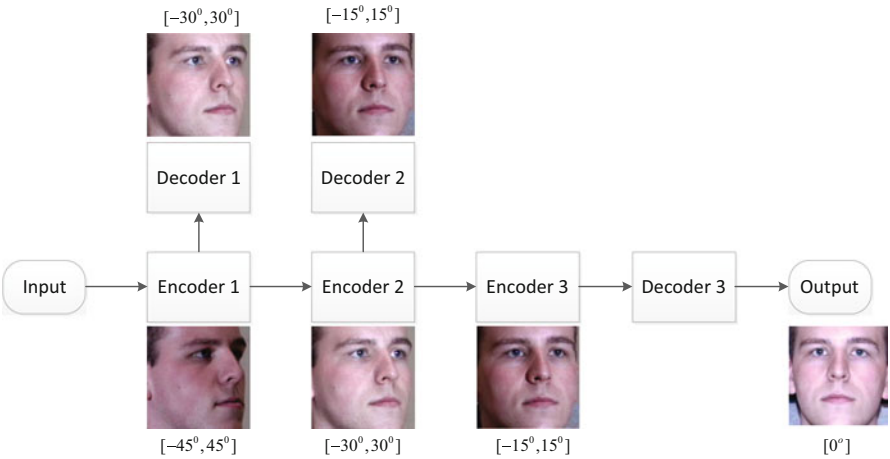

Fig. 9 Stacked progressive autoencoder network. The pose of the faces is adjusted step by step, and then all the trained encoders are stacked to compose a network to recover profile images to frontal ones

in Fig. 9. Altogether, three encoders are stacked together, which can reduce the probability of being trapped into local minimal during training.

For all these autoencoder-based methods, the key idea is to reconstruct the frontal face from non-frontal input. Then the recognition is performed based on the reconstructed images. Always, these methods separate the recognition tasks into two independent steps, which is not an end-to-end procedure.

Kan et al. [27] propose a two-stage network for face recognition across poses, as shown in Fig. 10. For the first stage, images from different views are input into different sub-networks for the extraction of view-specific features, and then all the features are fed into a common sub-network for the extraction of common features across poses. In addition, the network is trained based on the Fisher principle, where the intra-view distance is minimized and the inter-view discrepancy is maximized. With the trained network, the features from topmost layers are used for classification. For MvDN, it first requires the input images to be classified into groups due to poses, and then images can be fed into the proper network. In addition, it is not an end-to-end framework for the task of recognition.

Majumdar et al. [38] propose to use autoencoder for the extraction of image features. In order to make the feature more robust to pose, a whole face image is decomposed into several local patches that contain the main components. Also, the sparsity constraint is applied to the autoencoder. For classification tasks, the input images are fed to the autoencoder for feature extraction, and then a classifier is applied for verification. Even though the patch-based method can improve the robustness for pose, it requires accurate segmentation results from the preprocessing methods.

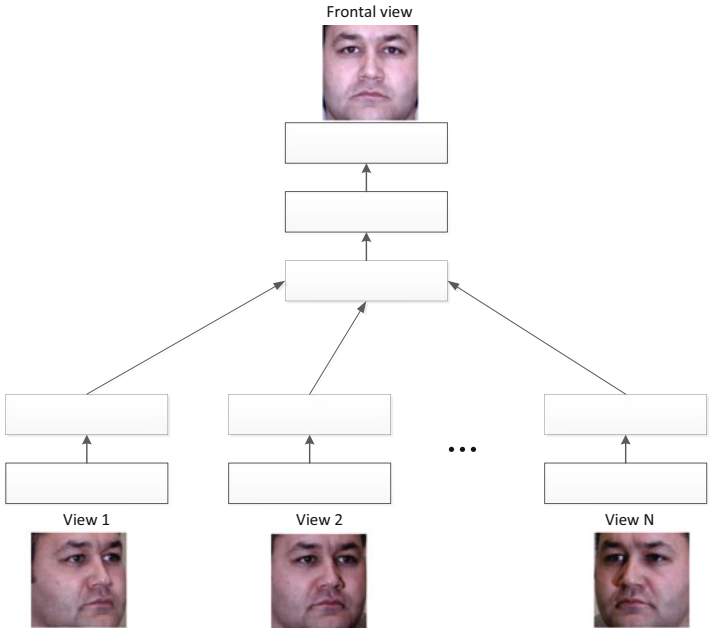

Fig. 10 Multiview deep network. The network first extracts features for different views separately, and then the features are integrated in the following network. In training the Fisher distance between the frontal view and different views is used as the error to optimize the parameters

Recently, Peng et al. [41] propose a deep network to extract pose-invariant features. First they use synthesis methods to enrich the training samples, and then identity and non-identity features are extracted through a multitask learning procedure. Finally, the pose-invariant features are purified through the constraint of reconstruction errors from different poses. The proposed method requires the synthesis of non-frontal faces for training. Also, it requires the pose and landmark labels for each training image, which is not easy to get in the applications.

Masi et al. [39] realized that the frontalization of non-frontal faces is actually very challenging and becomes harder with the increasing of rotation angle. Actually, it is a highly nonlinear transform, and many corresponding information between frontal and profile faces is lost. Thus they propose to develop separated network for different poses, called pose-aware CNN models (PAM). That is, for different poses, e.g., frontal, half-profile, and profile images, different CNNs are trained. Then averaging the scores from all different PAMs gives the recognition results. For this method, it requires the input images to be rendered into different poses, and then they can be fed into corresponding PAMs for recognition. Even though they constrain the range of rendered image, it is still a challenge to produce images of different views.

Tran et al. [61] introduce generative adversarial network (GAN) for the task of pose-invariant face recognition. The GAN network is constructed based on the

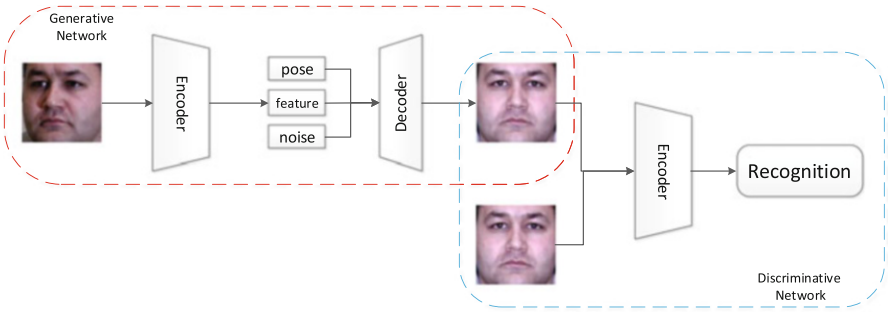

Fig. 11 Generative adversarial network (GAN)-based pose-invariant face recognition. In the generative network, the encoder-decoder structure is used. Also, some side information is introduced for the extraction of robust features in the network

framework of encoder and decoder. Through the generative network, face images from any poses are converted to frontal ones. The discriminative network is used for the judgment of the new created images. In order to enhance the extraction of pose-invariant feature in the generative network, pose label and noises are used as side information. The structure of the proposed network is shown in Fig. 11. Even though GAN network is more powerful than traditional neural network, the convergence of the network is still a great limitation for its application.

Neural network provides a powerful tool to extract features from training images, which has been applied to solve the pose problem. However, how to design a suitable network structure for this specific problem is still an open problem for researchers.

2.2 Synthesis-Based Methods

Pose variations introduce nonlinear transform for images of the same subject. Besides pose-invariant features that can be used as a low-dimensional representation for the image, researchers also tried to synthesis the frontal face images directly from images of arbitrary poses. The synthesis-based methods can be further classified into the 2D-based and 3D-based methods, depending on if the 3D model of face is applied or not.

2.2.1 2D-Based Synthesis Methods

2D-based synthesis methods try to convert images of varied poses directly to the frontal faces. Based on the different units to process, all the 2D synthesis methods can be classified into three categories: triangle mesh wrapping, patch wrapping, and pixel wrapping.

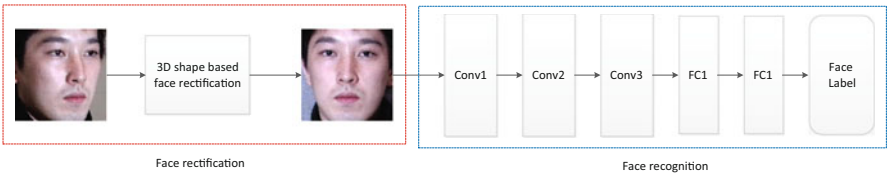

Fig. 12 Profile faces are first rectified to frontal face based on triangular mesh, and then deep network is applied to recognize faces in the frontal view

In the early days, a 3D model of a subject is usually presented by a triangle mesh. Consequently, the triangle mesh can also be casted on 2D images, and each triangle is used as a unit to calculate the transform between two different poses. Taigman et al. [59] apply triangle mesh for alignment of the input images. After all the input images are converted to frontal face, a deep network is applied for feature extraction and recognition of the face image, as shown in Fig. 12. For the alignment, key landmarks are detected and then triangular mesh is cast to the 2D image, where the 3D shape of the face is correlated with the triangular mesh. Finally the frontal facial image can be estimated due to the affine wrapping of triangular mesh. With accurate adaption of input images to frontal view, the following neural network can perform face recognition under varied poses. For mesh-based wrapping, it mainly depends on how well the triangle mesh can be cast to the 2D face image.

In order to avoid the detection of landmarks on face, Ashraf et al. [4] propose to decompose images into small patches, and for each patch, a learned transform can be applied to convert the non-frontal patch to be frontal ones. They treat each patch as independent unit for the whole procedure; however, there are close connections between neighboring patches. Thus those patches are highly correlated to each other where the relationship can be applied in solving the pose problem. Recently, Ho et al. [19] consider this relationship. The searching for optimal transform for each patches are converted into an optimization problem, where the reconstruction error of each patches should be small; meanwhile the transform for neighboring patches should be similar. The additional items constrain the smoothness of the global transforms.

For patch-based synthesis methods, they treat a local patch as a unit to calculate the transform. However, the nonlinear transform between poses is different from pixel to pixel. Thus Li et al. [34] propose a pixel-based transform. They learn a set of template displacement model from 3D dataset first. Then for each input image, the template displacement model is applied to transfer the images of arbitrary poses into frontal face pixel-wisely. For the occluded part of the face, the information is compensated from frontal face. Even though the proposed method can reconstruct the frontal face, only the non-occluded region will be used for verification.

2.2.2 3D-Based Synthesis Methods

For 2D-based wrapping methods, they directly find the transform between different poses. When the pose changes significantly, the mapping between different poses becomes highly nonlinear. As a result, the 2D wrapping results will become worse. In fact, pose problem is caused by the projection of 3D subject model to 2D imaginary surface. That is, the intrinsic 3D model controls the appearance of the images. Thus researches also try to use 3D model for solving pose-invariant recognition problem.

Ding et al. [12] introduce a 3D model-based dense-mapping method for the recovery of frontal face. First, the key landmarks are detected from the 2D images of arbitrary pose, and then they are matched to the corresponding landmarks in a standard 3D face model. As a result, the pose transform can be estimated for the input 2D image. In order to recover the texture, a dense mapping is used with the estimated pose transformation matrix. Furthermore, homography-based patch correction method is proposed to enhance the realisticity of the recovered texture. If there is occlusion in the original 2D facial images, then the recognition is only based on recovered un-occluded part.

Further, Ding et al. [13] transform the profiled face image recognition problem to the partial face recognition problem. Based on the key point mapping between 2D images and 3D face model, the profiled images are transformed to images of frontal view. Then sparse coding-based feature is extracted on the reliable regions of the recovered frontal view.

3 Illumination-Invariant Face Recognition

Illumination is another big challenge for face recognition. The intensity value of each pixel I_{xy} in an image is determined by the strength of the incident light, L_{xy}, and the angle of the incident light θ_{xy} and the reflectance rate of the surface R_{xy}, as $I_{xy} = \oint L_{xy} R_{xy} \cos\theta_{xy} d\Omega$. Thus, when the incident light changes, the appearance of the same object will vary as well. Always, the variation that is caused by illumination is more significant than that of subject. Consequently, lighting always causes the degradation of face recognition methods. Algorithms that try to remove or alleviate the lighting variations can be classified into three categories as image processing-based methods, invariant feature-based methods, and illumination model-based methods.

3.1 Image Processing-Based Methods

Lighting is one of the factors that control the appearance of images. The average intensity value of images that are taken in brighter situation is bigger than that

in dark situation. Thus researches proposed to use image-processing methods to enhance the intensity value of images, such as histogram equalization (HE) and gamma correction.

Histogram equalization [43] tries to adjust the histogram of input images. That is to adjust the intensity distributions of images. In fact, the pixel intensity value of an image that is taken in brighter situations is bigger than that of a darker image. Thus the intensity distribution of brighter images will have peaks in bigger value region, while that of darker images will have peaks in smaller value region. Using histogram equalization will make the distribution of intensity value evenly. That is, the brighter images will become dimmer, and darker ones will become brighter. Histogram equalization only considers the intensity value of each pixel but not the sematic meaning. Therefore it will always introduce some abrupt noises in images due to the assignment of new intensity value to all the pixels of the same intensity value, as shown in Fig. 13b.

Gamma correction is the more dedicated adjustment for the intensity distribution, while histogram equalization turns the original distributions into uniform distribution. Gamma correction function is defined as

$$T(I) = I_{\max}\left(\frac{I}{I_{\max}}\right)^{\gamma} \tag{5}$$

where I is the intensity value of current pixel. I_{max} is the maximum intensity value in the image, and γ determines the curve for adaption, which is the key parameter. According to different gamma curves, the original intensity values can be modified to any desired distributions. Normally the darker pixels are tuned to be brighter, while the bright pixels are kept, as shown in Fig. 13c.

Huang et al. [21] proposed an adaptive gamma correction method, where γ is determined by the cumulative distribution of intensity, $cdf(I)$, as

$$T(I) = I_{\max}\left(\frac{I}{I_{\max}}\right)^{1-cdf(I)} \tag{6}$$

$$cdf(I) = \sum_{I=0}^{I_{\max}} \frac{pdf(I)}{\sum pdf} \tag{7}$$

where $pdf(I)$ is the probability density function of intensity. The proposed method combines gamma correction and histogram modification method. Similar to the histogram equalization, gamma correction is also a holistic modification method, which does not consider the local information in the image. Jiang et al. [23] proposed to combine the local and global information for the lighting augment. The local factor I_a^{local} provides the local variance, while the global factor I_a^{global} provides the overall intensity of the image. These two kinds of information are combined by a bilinear method, and then a perception-based method is used to adjust the brightness of the images, as

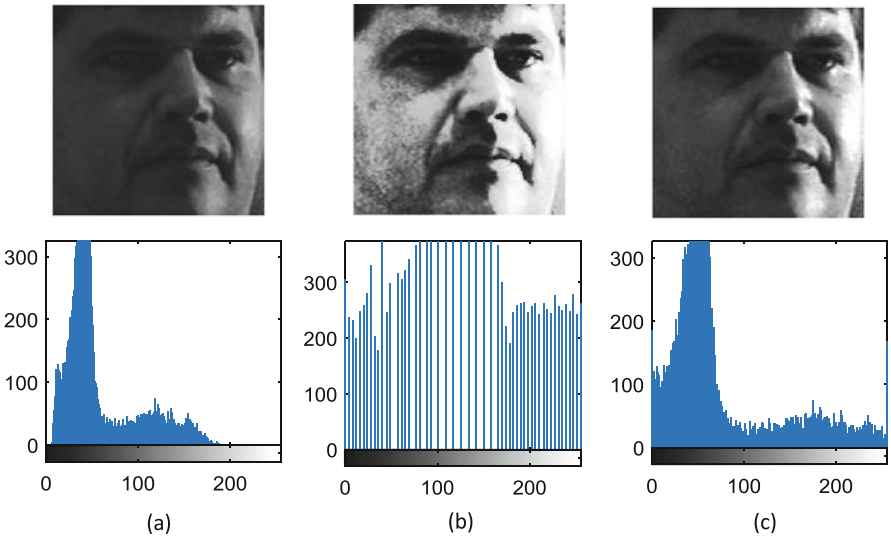

Fig. 13 Comparison of two basic illumination adjustment methods. (**a**) The original image and its histogram. There are two peaks in the histogram due to the side light. (**b**) Histogram equalization result. The histogram is adjusted to be equally distributed but leaves many noises in the image. (**c**) Gamma correction result. Compared with histogram equalization result, there is few noise introduced into the result image

$$Y(\alpha, m, f; I) = \frac{I}{I + (f I_a)^m} I_{\max} \tag{8}$$

$$I_a = \alpha I_a^{\text{local}} + (1 - \alpha) I_a^{\text{global}} \tag{9}$$

where α, m, and f are parameters determine the detail, contrast, and brightness of the image. The model is derived from human vision perception system. Image processing-based illumination adjustment methods mainly focus on the modification of the intensity distribution and aim to brighten dimmed images. This kind of method only considers the appearance of current images, but not the factors that cause the current appearance. Thus these methods always cannot solve the lighting problems thoroughly.

3.2 Invariant Feature-Based Methods

Images are the cooperative results of illumination and objects. Even though the illumination can vary due to different situations, objects themselves do not change. Therefore researchers try to find illumination-invariant representations from images to describe the intrinsic features of objects. Edges describe the shape or contour

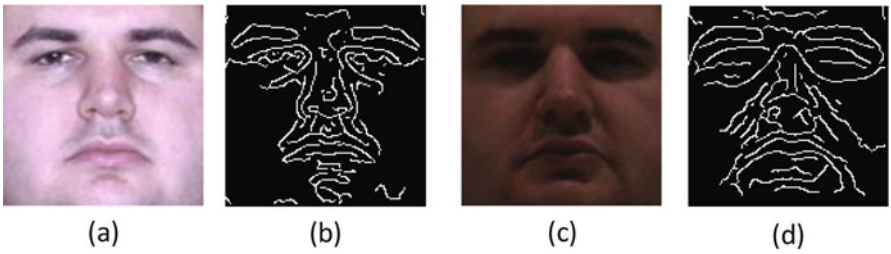

Fig. 14 Edges are used as features for face recognition. (**a**) and (**c**) are the original images, (**b**) and (**d**) are the edges for (**a**) and (**c**), respectively. Compared to (**b**) and (**d**), the edges are different for the same person under different lighting conditions. Thus the edge features are not absolutely invariant to illumination

of objects, which are considered as one of the illumination-invariant features. On the contrary, color of the object will vary according to its situation. Gao et al. [15] propose to use edges on the face image to perform face recognition. All the face components, such as eyes, nose, mouse, and eyebrows, are represented by line contours, as shown in Fig. 14. Then the distance between lines is calculated for verification. Zhou et al. [72] apply multi-scale Gabor filters to extract features from face images, where the multi-scale edge features are extracted. Some other popular feature descriptors such as local binary pattern (LBP) and scale-invariant feature transform (SIFT), which extract edge-based features for local regions, are also considered to be robust to illumination. However, shadow will also produce edges, even obvious edges in images, which are quite difficult to be distinguished from edges of the object. Consequently, these edge-based features can only work well with slight lighting variations.

With the simplified illumination model $I = L \times R$, Shashua et al. [56] propose the concept of quotient images, which is the ratio between a testing image I_y and a linear combination of three images I_j with weight x_j, as

$$Q_y = \frac{I_y}{\sum x_j I_j} \tag{10}$$

where the combined lighting condition of I_j is similar to the lighting condition of I_y. Thus the quotient image only relates to the reflectance of the object and is free from the lighting variations. With quotient image Q_y, images under new lighting condition can be rendered and furthermore can be used for the face recognition in different lighting conditions. However, quotient image is based on the assumption that the same class of object all has the same shape. It is a very rough assumption. In fact, every face is different. Wang et al. [64] extend the concept of quotient images. They propose to estimate the lighting map from images directly. According to the Retinex theory [60], most lighting information can be considered as the low-frequency signal, and most reflectance information is high frequency. Thus lighting information can be estimated from the low-frequency part of the original images. The proposed self-quotient image is defined as

$$Q_{sy} = \frac{I_y}{F * I_y} \tag{11}$$

where F is a smoothing kernel.

Also based on the theory of Retinex, Xie et al. [68] propose a two-step strategy for the normalization of lighting conditions. First, they decompose the input images into the low-frequency and high-frequency parts using the total variation model. Then the two decomposed components are normalized, respectively. The normalization of the low-frequency part will enhance the uniformity of lighting conditions. Then the normalized high-frequency and low-frequency parts are multiplied together to get the normalized images. In the second step, kernel eigenspace is used to correct the visual flaws of the normalized face images. Even though KPCA can be used to improve the appearance of the image, it requires a lot of training images of each subject for the construction of kernel subspace.

He et al. [18] realize that the distribution of face subspace is a nonlinear manifold; thus, the nonlinear method should be more suitable for the problem. They propose to find the face manifold based on the locality preserving projection, which is called as the Laplacian face representation. This subspace can keep the identity difference but minimize the other variance within a same subject.

Compressive sensing theory provides a dramatically new method to represent signals. Based on the theory, continuous signals can be sampled randomly which breaks through the constraint of Shannon theorem. From the training dataset, a sparse representation of the subject can be learned with the sparse constraint. Wagner et al. [63] construct a sparse coding dictionary from a set of images taken in different lighting conditions and different poses. Images in various conditions are recovered to classical frontal images of the same subject with the sparsity constraint. Then the learned sparse representation is invariant to illumination. However, the sparse dictionary learning requires images from all different conditions to keep the performance of the proposed algorithm.

Recently, deep learning-based methods are also applied to extract illumination-invariant features for images under different lighting conditions. The classical deep learning methods, such as AlexNet [29] and VGG-Face network [40], extract features from the convolutional layers, and then discriminative features are classified by the fully connected layers. Besides the classical neural network structures, researchers also adapt loss functions to improve the extracted features [20, 66]. These methods do not focus on the lighting problems but try to extract robust features for general face recognition problem. Therefore the structure of the network is not specifically designed for lighting problem.

3.3 Illumination Model-Based Method

Illumination is an essential factor for imaging. Therefore, researchers also try to analyze the distributions of images that are taken in different lighting conditions.

With the illumination model, images under different lighting conditions can be reconstructed. Also, the illumination can be changed or removed from the images. Belhumeur and Kriegman [7] introduce the theory of illumination cone, which is the basic theory for the lighting space of an object. If an object has convex shape and Lambertian surface, then all the images about this object can form a polyhedral cone. The dimension of this cone is determined by the number of distinct surface normal vectors of the subject. In practice, the illumination cone theory can be relaxed to objects of any shape and with a general reflectance surface. The authors also point out that the illumination cone of an object could be approximated by a low-dimensional subspace. The illumination cone theory only illustrates the structure of lighting space for an object in a certain pose. The relationship between illumination cones for different poses is still unclear.

In fact, a completed high-dimensional illumination cone is always difficult to build in practice. Thus, researchers try to find the low-dimensional approximation for an illumination cone. The illumination effect for an object can be considered as the convolution of incident illumination with the reflectance function of the object. Given the 3D model of an object, the lighting subspace can be constructed by spherical harmonic basis [5, 45]. All the bases are given with implicit equations, which are functions about illumination and object surface normal vectors. Given the 3D shape of the object, lighting position, and intensity, it is very easy to obtain the basis for the lighting subspace directly. With the first three orders of the harmonic basis, 90% of the illumination effect can be estimated, as shown in Fig. 15. However, the requirement of deep information for the object also limits the application of the method.

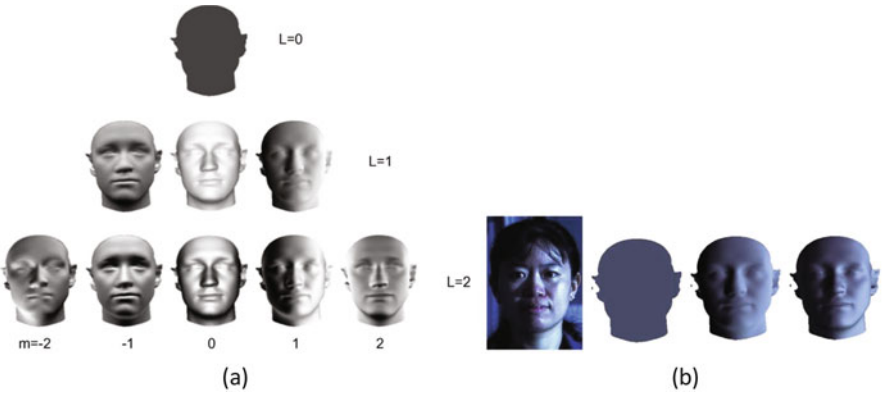

Fig. 15 Lighting map estimation based on spherical harmonics basis. (**a**) the first three orders of the spherical harmonic basis of a face. Each row is the basis of the same order. From top to bottom, they are the basis of order 0 to order 2, respectively. (**b**) From left to right, they are the original face image, lighting map reconstructed by the basis of order 0 to 2, respectively. (Reprinted from Ref. [22], with permission from Elsevier)

Fig. 16 Images taken in nine specific lighting conditions can be used as the basis to construct the lighting subspace of the object. (Reprinted from Ref. [24], with permission of Springer)

In order to avoid the requirement of 3D model of subject or the learning procedure to find lighting subspace, Lee et al. [30] propose to construct the low-dimensional approximation directly from real images. Those real images are found through minimizing the distance between two subspaces, where one is the spherical harmonic subspace H and the other is the selected real image subspace, C. Then the real lighting configuration of those selected images can be used for any subject to construct the low-dimensional lighting subspace, C. That is, the real images taken under those lighting configuration can be used as the basis images for its lighting subspace. This paper proves that the lighting subspace constructed from real images is a very good approximation of spherical harmonic subspace. In application, only nine lighting positions are required to build up the lighting subspace, as shown in Fig. 16. Also, it does not require the depth information as the traditional spherical harmonic subspace. However, the specific lighting configuration sometimes cannot be accessible. In real application, there are always a few sample images or even one image of each subject.

Considering the practical application of lighting subspace estimation, Jiang et al. [24] propose to create the basis images from any sample images of an object. That is, given one image that can be taken in arbitrary condition, the nine basis images can be reconstructed, and consequently the lighting subspace can be set up, as shown in Fig. 17. In this paper, the estimation of basis images is based on the maximum a posterior estimation. The basis images are composed of the common components and personal components, where the common component is the mean value from the training images and the personal components describe the specific characteristics of each subject. Thus the estimated basis images can recover both the shared and the individual features for each subject. This method breaks the requirement of nine real images that are taken in specific lighting condition.

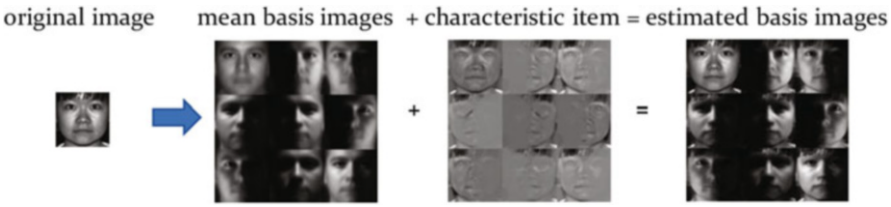

Fig. 17 The basis images of lighting subspace can be estimated from images under arbitrary lighting conditions

4 Multi-stream Convolutional Neural Networks

In general conditions, the illumination of the environment and the pose of the object are always uncontrolled. Therefore the robust face recognition system should be able to process the pose and lighting problems at the same time. The current algorithms that are dealt with pose and lighting problems are introduced in Sects. 2 and 3, respectively. Besides these specific designed algorithms, there are also some methods doing face recognition in general conditions. Especially with the development of deep learning methods, some deep neural networks are designed for the face recognition problems. Schroff et al. [50] propose to enhance the discrimination of the deep features according to the standard that the distance within a class is minimized and the distance between classes are maximized. Sun et al. [57] increase the dimension of the hidden layers and add constraint for early convolutional layer to increase the discriminative power of the neural network. The proposed network is called as DeepID2+, which improve the performance for the face recognition in natural conditions.

For face recognition across pose and illumination, the global structure of images is destructed by views; meanwhile, lighting brings wide variations for the appearance of images. Thus the pose- and illumination-invariant features should be local but not global. Furthermore the multiple hierarchical features are always much more informative than features in a single scale. Consequently, we propose an end-to-end convolutional network which can extract multi-hierarchy local features for the task of face recognition. The overall architecture is shown in Fig. 18. In our proposed networks, the input is a facial image under an arbitrary pose and illumination. The output is the identity label for the face image.

4.1 Root Convolutional Layer

Recently convolutional neural networks (CNN) show great performance in different fields of computer vision, such as object detection [46] and object classification or recognition [58]. The superb capability of CNN is mainly due to its high-order

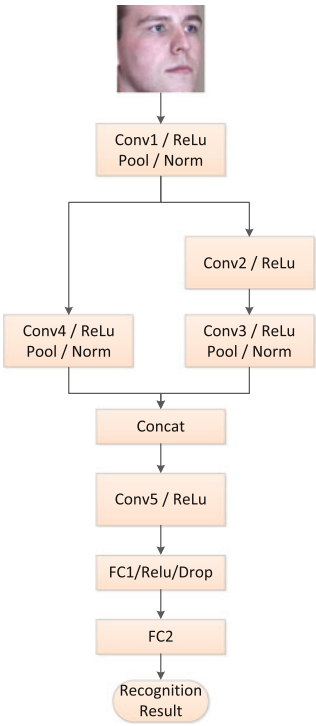

Fig. 18 Architecture of the proposed deep network. Conv1 has the kernel size of 11×11 and the dimension of 96. Conv2, Conv3, Conv4, and Conv5 all have the kernel size of 1×1 and the dimensions of 200, 400, 300, and 500, respectively

nonlinear representation for data. In practice, CNN extracts features from input images layer by layer through convolution kernels. For the proposed networks, the input images, x_i^{l-1}, are cropped and mirrored to the size of $w \times h \times c = 227 \times 227 \times 3$. Then they are fed into a convolutional layer (Conv1) k_{ij}^l with 96 filters of size $11 \times 11 \times 3$. The output, x_j^l, of this convolutional layer is written as

$$x_j^l = \sum_{i \in M_j} x_i^{l-1} * k_{ij}^l + b_j^l \tag{12}$$

where l is the layer index, b_j^l is the additive bias term, and $*$ represents convolution in a local region M_j of input signals. In this convolutional layer, 96 filters are applied locally to the whole images resulting in a feature map of size $55 \times 55 \times 96$. Then rectified linear unit (ReLU) is applied to the extracted feature map. ReLU serves as an activation unit in the network which brings the nonlinearity to the feature. Here we use a ramp function $f(x) = max(0, x)$ to rectify the feature map. This activation function is considered to be more biologically plausible than the widely used logistic sigmoid or hyperbolic tangent function.

Consequently, the rectified feature maps will be given to a max-pooling layer (Pool) which takes the max over 3×3 spatial neighborhoods with a stride of two for

each channel, respectively. Through max-pooling operation, features will become insensitive to local shift, i.e., invariant to location. Afterward, those features will go through the local response normalization layer (Norm), which performs the lateral inhibition by normalizing over local input regions. In our model, the local regions extend across nearby channels but have no spatial extent. For normalization, each input value is divided by the sum of local region as shown in Eq. 13:

$$s(x_i) = (k + (\alpha/n) \sum_i x_i^2)^\beta \tag{13}$$

where n is the size of each local region and the sum is taken over the region centered at that value x_i (zero padding is added where necessary). From the root layer, we can obtain the local feature set which mainly contains all kinds of edges in different orientations. Generally edges are considered as illumination-invariant features. Since local structures are more important in our case, we will continue to seek for local features instead of global features which are normally obtained in further layers of traditional CNN.

4.2 Multi-hierarchical Local Feature

In fact, the window size of the convolution kernel is considered as the receptive field for feature extraction. That is, bigger windows can include information in wider range for processing. For the face images from different views, the global structures of images change diversely. However, there is a tight correlation among local regions of images taken in different views. Therefore the pose-invariant features should be local features, and in addition the spatial information should be kept for each local feature. Accordingly, smaller windows should be applied to extract features. Here we propose to use the kernel of size $1 \times 1 \times c$. With the kernel of 1×1, no spatial patterns across multiple pixels are extract, but the patterns between c channels are learned without losing the location information for each pattern. Thus the feature can keep the correlation among different views. Meanwhile, the number of parameters will also be reduced for the 1×1 kernel size compared with that of bigger kernel size, which can make the training procedure more easily to be convergent.

Neural networks actually perform nonlinear operation for data. With multiple layers of processing, the neural networks can build a high-order nonlinear model for real images, which is a much more suitable representation for the data than the traditional man-crafted features. As a result, different numbers of layers also influence the property of features. In the classical ConvNet, there is only one path for the signal to go, and the classification only performs on the features extracted by the last layer of the network. In fact, features from different levels of the networks all contain useful information. Thus we propose to build a multi-stream local feature hierarchy network (LFHN). Within each stream, features of different orders are

extracted by using different numbers of convolutional layers. Then features from different streams are contact to compose a multi-hierarchy feature with the size of $h \times w \times (c_1 + c_2 + \cdots + c_n)$, where $h \times w \times c_i$ is the size of the feature from stream i and c_i is the depth dimension.

In order to keep the spatial information for the local features, the convolutional kernel of 1×1 is applied. As shown in Fig. 18, there are two steams for the proposed network. One stream contains two convolutional layers where the kernel size of Conv2 and Conv3 is $1 \times 1 \times 200$ and $1 \times 1 \times 400$, respectively. In another stream, there is only one convolutional layer Conv4 with the kernel of $1 \times 1 \times 300$. Through these two streams, we can get local features from different levels of hierarchy. In order to achieve the final recognition task, one more convolution layer, Conv5 with kernel size $1 \times 1 \times 500$, and two fully connected layer are employed for the further feature abstraction.

4.3 Training

Since the root layers of convolutional networks always contain more generic features such as edges or color blob, which is useful for many tasks including face recognition, in training, we keep the pretrained results of AlexNet as the weight for the convolutional layer, Conv1. Then for the other layers, they are trained according to the Softmax loss function based on the identity labels for images in MultiPIE dataset [16].

5 Experiments

5.1 Dataset

To evaluate the effectiveness of the proposed local feature hierarchy networks (LFHN) under different poses and illumination, the MultiPIE face database [16] is employed. The MultiPIE face database contains 754,204 images of 337 identities. Each identity has images captured under 15 different poses and 20 different lighting conditions. For the original images in MultiPIE, we have aligned all the images according to the position of eyes and crop them to the size of 256×256. For each subject, we only select the images with neural expression but in all poses and lighting conditions; thus for each person, there are 300 images. Altogether, for all the individuals in the dataset, we put $300 \times 337 = 101{,}100$ images into the data pool for training and testing. In MultiPIE, there are four sessions to take photos for each subject, but not everyone comes in each session. Therefore we take all the images of 250 individuals in session 1 and the images of the other 87 persons in session 2.

5.2 *Recognition Across Poses and Illumination*

In order to evaluate the performance of the proposed local hierarchy networks, MultiPIE is used to train and test the networks. For the proposed networks, there is only one input instead of multiple images from different reviews. Therefore training images are randomly selected from images of natural expression but in 15 different poses and 20 different lighting conditions. For all the 337 individuals, 90,000 images are randomly taken from 101,100 images for training, and the leftover images are used as testing images. The proposed network can learn local nonlinear features which can represent the correlations between images in different poses and lighting conditions. The rank-1 recognition rates for images with pose and illumination variations are shown in Table 1. The recognition results for each view are the average results for all the images under 20 different lighting conditions.

From the results, we can see that the proposed LFHN network achieved relatively stable performance for different poses. Especially for profile-wised images, where the yaw angle is in the range of $[-90°, -60°]$ and $[60°, 90°]$, the average recognition rate is 97.78%, while for the traditional methods, the performance declined significantly for images with greater pose variations. For the patch-based partial recognition (PBPR) [14], the average recognition rate for front-wised images is 98.96% where the yaw angle is within 45°. But for the profile-wised images, the recognition rate is 78.76%. That indicates the projection recovery method used in PBPR does not find the accurate locations for profile-wised images. Compared with current state-of-the-art methods, the proposed LFHN network improves the recognition rate by 7.55% for images under arbitrary poses and illumination. In the proposed networks, we consider images from different views and illumination

Table 1 Rank-1 identification rates on combined variations of pose and illumination on MultiPIE

PoseID	Yaw	Pitch	RR [32]	FIP [73]	PBPR [14]	LFHN (Ours)
081	−45°	25°	24	–	88	**94.51**
110	−90°	0°	20.5	–	51	**97.52**
120	−75°	0°	26.5	–	79	**98.15**
090	−60°	0°	50.64	–	90.86	**98.51**
080	−45°	0°	65.30	67.10	**97.91**	97.75
130	−30°	0°	70.97	74.60	**99.41**	98.08
140	−15°	0°	81.07	86.10	**99.05**	97.12
050	15°	0°	77.21	83.30	**99.94**	93.74
041	30°	0°	73.69	75.30	**99.23**	97.91
190	45°	0°	58.12	61.80	**98.21**	96.53
200	60°	0°	45.97	–	87.75	**97.65**
010	75°	0°	31	–	89	**97.57**
240	90°	0°	18	–	75	**97.3**
191	45°	25°	40	–	96	**93.73**
		Mean	48.78	–	89.31	**96.86**

Table 2 Rank-1 identification rates for pose variation on MultiPIE

PoseID	Pose	PLS [54]	MCCA [48]	PLS + LDA [27]	MCCA + LDA [27]	MvDA [28]	GMA [52]	MvDN [27]	LFHN (ours)
110	−90°	0.319	0.409	0.38	0.488	0.568	0.526	0.704	**1**
120	−75°	0.775	0.742	0.798	0.662	0.723	0.723	0.822	**0.9767**
090	−60°	0.892	0.822	0.869	0.817	0.845	0.845	0.883	**1**
080	−45°	0.934	0.723	0.944	0.887	0.92	0.901	0.911	**1**
130	−30°	0.883	0.685	0.92	1	0.967	1	**0.991**	0.893
140	−15°	0.981	0.92	0.995	1	1	1	**1**	1
050	15°	0.981	0.906	0.986	1	1	1	**1**	0.938
041	30°	0.934	0.798	0.967	0.995	0.991	1	**0.991**	0.971
190	45°	0.906	0.747	0.883	0.831	0.897	0.906	0.93	**0.958**
200	60°	0.873	0.779	0.85	0.803	0.864	0.859	0.911	**0.935**
010	75°	0.723	0.714	0.709	0.676	0.714	0.718	0.798	**1**
240	90°	0.268	0.376	0.319	0.568	0.559	0.573	0.709	**1**
	Mean	0.789	0.718	0.802	0.811	0.837	0.838	0.887	**0.973**

equally. Thus pose-invariant and illumination-invariant nonlinear local features can be sought by the proposed network LFHN.

Besides the face recognition with the combined variations of pose and illumination, we also test the performance of the proposed network LFHN on pose only. For this task, the probe dataset includes images of all subjects from four sessions where images are taken in ambient lighting and 13 different poses. The comparison results with other methods are shown in Table 2.

From the results we can see even though different methods try to tackle the pose problem, the face recognition rate decreases along with the amount of pose variation. The more the view diverges from the frontal face, the lower the recognition rate is. That is because pose changes the appearance and structure of the image. MvDN [27] and MvDA [28] tried to find the correlation between different poses and achieved relatively good performance. For the proposed network LFHN, we also focus on the extraction of local features among different poses which can describe the correlation of different poses and also discriminate different identity. Thus we achieve better results compared with other methods. Especially for larger pose diversity, the performance of the proposed network is not degenerated but very stable instead. The average recognition rate is 97.3% which improves the state-of-the-art method by 8.6%.

6 Conclusion

Pose and illumination will always bring great variance for the appearance of face images, which makes face recognition across pose and illumination challenged. However, it is quite normal to encounter the pose and illumination changes in uncontrolled environment. Therefore a robust face recognition system has to deal

with illumination and pose variations effectively. In fact, there are tight correlations for images from different postures. Images of different views are the projection of the same object to different positions. Then local features are more useful for recognition under different views than the global features, where the global structure is actually destructed by the projection in different views. Thus we propose a neural network which extracts local features by 1×1 convolutional kernels; in addition multi-hierarchical features are combined for the task of recognition. Experiments on MultiPIE dataset show very good and stable performance for the proposed networks in a wide range of pose and illumination.

Acknowledgements This chapter is partly supported by the National Natural Science Foundation of China (No.61502388), Ph.D. Programs Foundation of Ministry of Education of China (No. 20136102120041), the Fundamental Research Funds for the Central Universities (No. 3102015BJ (II)ZS016), and the Shaanxi Province International Science and Technology Cooperation and Exchange Program (2017KW002).

References

1. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **28**(12), 2037–2041 (2006). DOI 10.1109/TPAMI.2006.244
2. Ahonen, T., Hadid, A., Pietikainen, M.: Face description with local binary patterns: Application to face recognition. IEEE transactions on pattern analysis and machine intelligence **28**(12), 2037–2041 (2006)
3. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: International Conference on Machine Learning, pp. 1247–1255 (2013)
4. Ashraf, A.B., Lucey, S., Chen, T.: Learning patch correspondences for improved viewpoint invariant face recognition. In: Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pp. 1–8. IEEE (2008)
5. Basri, R., Jacobs, D.W.: Lambertian reflectance and linear subspaces. IEEE Transactions on Pattern Analysis and Machine Intelligence **25**(2), 218–233 (2003)
6. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-up robust features (surf). Computer vision and image understanding **110**(3), 346–359 (2008)
7. Belhumeur, P.N., Kriegman, D.J.: What is the set of images of an object under all possible illumination conditions? International Journal of Computer Vision **28**(3), 245–260 (1998)
8. Biswas, S., Aggarwal, G., Flynn, P.J., Bowyer, K.W.: Pose-robust recognition of low-resolution face images. IEEE transactions on pattern analysis and machine intelligence **35**(12), 3037–3049 (2013)
9. Castillo, C.D., Jacobs, D.W.: Using stereo matching with general epipolar geometry for 2d face recognition across pose. IEEE Transactions on Pattern Analysis and Machine Intelligence **31**(12), 2298–2304 (2009)
10. 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, pp. 3025–3032 (2013)
11. Ding, C., Choi, J., Tao, D., Davis, L.S.: Multi-directional multi-level dual-cross patterns for robust face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence **38**(3), 518–531 (2016). DOI 10.1109/TPAMI.2015.2462338
12. Ding, C., Tao, D.: Pose-invariant face recognition with homography-based normalization. Pattern Recognition **66**, 144–152 (2017)

13. Ding, C., Xu, C., Tao, D.: Multi-task pose-invariant face recognition. IEEE Transactions on Image Processing **24**(3), 980–993 (2015)
14. Ding, C., Xu, C., Tao, D.: Multi-task pose-invariant face recognition. IEEE Transactions on Image Processing **24**(3), 980–93 (2015)
15. Gao, Y., Leung, M.K.: Face recognition using line edge map. IEEE transactions on pattern analysis and machine intelligence **24**(6), 764–779 (2002)
16. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. Image and Vision Computing **28**(5), 807–813 (2010)
17. Gunther, M., Costa-Pazo, A., Ding, C., Boutellaa, E.: The 2013 face recognition evaluation in mobile environment. 2013 International Conference on Biometrics (ICB) pp. 1–7 (2013)
18. He, X., Yan, S., Hu, Y., Niyogi, P., Zhang, H.J.: Face recognition using laplacian faces. IEEE transactions on pattern analysis and machine intelligence **27**(3), 328–340 (2005)
19. Ho, H.T., Chellappa, R.: Pose-invariant face recognition using markov random fields. IEEE transactions on image processing **22**(4), 1573–1584 (2013)
20. Hu, J., Lu, J., Tan, Y.P.: Discriminative deep metric learning for face verification in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1875–1882 (2014)
21. Huang, S.C., Cheng, F.C., Chiu, Y.S.: Efficient contrast enhancement using adaptive gamma correction with weighting distribution. IEEE Transactions on Image Processing **22**(3), 1032–1041 (2013)
22. Jiang, X., Cheng, Y., Xiao, R., Li, Y., Zhao, R.: Spherical harmonic based linear face de-lighting and compensation. Applied Mathematics and Computation **185**(2), 857–868 (2007). https://doi.org/10.1016/j.amc.2006.06.090. http://www.sciencedirect.com/science/article/pii/S0096300306007673. Special Issue on Intelligent Computing Theory and Methodology
23. Jiang, X., Feng, X., Wu, J., Peng, J.: Lighting alignment for image sequences. In: International Conference on Image and Graphics, pp. 462–474. Springer (2015)
24. Jiang, X., Kong, Y.O., Huang, J., Zhao, R., Zhang, Y.: Learning from real images to model lighting variations for face images. In: European Conference on Computer Vision (ECCV), pp. 284–297 (2008)
25. Kafai, M., An, L., Bhanu, B.: Reference face graph for face recognition. IEEE Transactions on Information Forensics and Security **9**(12), 2132–2143 (2014)
26. Kan, M., Shan, S., Chang, H., Chen, X.: Stacked progressive auto-encoders (spae) for face recognition across poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 1883–1890 (2014)
27. Kan, M., Shan, S., Xilin., C.: Multi-view deep network for cross-view classification. In: IEEE Conference on Computer Vision and Pattern Recognition (2016)
28. Kan, M., Shan, S., Zhang, H., Lao, S., Chen, X.: Multi-view discriminant analysis. IEEE transactions on pattern analysis and machine intelligence **38**(1), 188–194 (2016)
29. 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)
30. Lee, K.C., Ho, J., Kriegman, D.: Acquiring linear subspaces for face recognition under variable lighting. IEEE Transactions on Pattern Analysis and Machine Intelligence **27**(5), 684–698 (2005)
31. Li, A., Shan, S., Chen, X., Gao, W.: Maximizing intra-individual correlations for face recognition across pose differences. In: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on, pp. 605–611. IEEE (2009)
32. Li, A., Shan, S., Gao, W.: Coupled bias-variance tradeoff for cross-pose face recognition. IEEE Transactions on Image Processing **21**(1), 305–15 (2012)
33. Li, H., Hua, G., Lin, Z., Brandt, J., Yang, J.: Probabilistic elastic matching for pose variant face verification. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3499–3506 (2013)
34. Li, S., Liu, X., Chai, X., Zhang, H., Lao, S., Shan, S.: Morphable displacement field based image matching for face recognition across pose. European Conference on Computer Vision 2012 pp. 102–115 (2012)

35. Liao, Q., Leibo, J.Z., Poggio, T.: Learning invariant representations and applications to face verification. In: Advances in Neural Information Processing Systems, pp. 3057–3065 (2013)
36. Liao, S., Jain, A.K., Li, S.Z.: Partial face recognition: Alignment-free approach. IEEE Transactions on pattern analysis and machine intelligence **35**(5), 1193–1205 (2013)
37. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International journal of computer vision **60**(2), 91–110 (2004)
38. Majumdar, A., Singh, R., Vatsa, M.: Face recognition via class sparsity based supervised encoding. IEEE transactions on pattern analysis and machine intelligence (2016)
39. Masi, I., Rawls, S., Medioni, G., Natarajan, P.: Pose-aware face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4838–4846 (2016)
40. Parkhi, O.M., Vedaldi, A., Zisserman, A., et al.: Deep face recognition. In: British Machine Vision Conference, vol. 1, p. 6 (2015)
41. Peng, X., Yu, X., Sohn, K., Metaxas, D., Chandraker, M.: Reconstruction for feature disentanglement in pose-invariant face recognition. arXiv preprint arXiv:1702.03041 (2017)
42. Pentland, A., Moghaddam, B., Starner, T., et al.: View-based and modular eigenspaces for face recognition. In: CVPR, vol. 94, pp. 84–91 (1994)
43. Pizer, S.M., Amburn, E.P., Austin, J.D., Cromartie, R., Geselowitz, A., Greer, T., ter Haar Romeny, B., Zimmerman, J.B., Zuiderveld, K.: Adaptive histogram equalization and its variations. Computer vision, graphics, and image processing **39**(3), 355–368 (1987)
44. Prince, S.J., Elder, J.H., Warrell, J., Felisberti, F.M.: Tied factor analysis for face recognition across large pose differences. IEEE Transactions on pattern analysis and machine intelligence **30**(6), 970–984 (2008)
45. Ramamoorthi, R., Hanrahan, P.: On the relationship between radiance and irradiance: determining the illumination from images of a convex lambertian object. Journal of the Optical Society of America, A **18**(10), 2448–2459 (2001)
46. Ren, S., He, K., Girshick, R., Sun, J.: Faster r-cnn: Towards real-time object detection with region proposal networks. Advances in Neural Information Processing Systems (NIPS) pp. 91–99 (2015)
47. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. In: Conference on Data Mining and Data Warehouses (SiKDD 2010), pp. 1–4 (2010)
48. Rupnik, J., Shawe-Taylor, J.: Multi-view canonical correlation analysis. Conference on Data Mining and Data Warehouses(SiKDD 2010) pp. 1–4 (2010)
49. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. Neural computation **10**(5), 1299–1319 (1998)
50. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. Computer Vision and Pattern Recognition pp. 815–823 (2015)
51. Schroff, F., Treibitz, T., Kriegman, D., Belongie, S.: Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison. In: Computer Vision (ICCV), 2011 IEEE International Conference on, pp. 2494–2501. IEEE (2011)
52. Sharma, A.: Generalized multiview analysis: A discriminative latent space. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 2160–2167 (2012)
53. Sharma, A., Al Haj, M., Choi, J., Davis, L.S., Jacobs, D.W.: Robust pose invariant face recognition using coupled latent space discriminant analysis. Computer Vision and Image Understanding **116**(11), 1095–1110 (2012)
54. Sharma, A., Jacobs, D.W.: Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 593–600 (2011)
55. Sharma, A., Kumar, A., Daume, H., Jacobs, D.W.: Generalized multiview analysis: A discriminative latent space. In: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, pp. 2160–2167. IEEE (2012)
56. Shashua, A., Riklin-Raviv, T.: The quotient image: Class-based re-rendering and recognition with varying illuminations. IEEE Transactions on Pattern Analysis and Machine Intelligence **23**(2), 129–139 (2001)

57. Sun, Y., Wang, X., Tang, X.: Deeply learned face representations are sparse, selective, and robust. Computer Vision and Pattern Recognition
58. Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., Rabinovich, A.: Going deeper with convolutions. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
59. Taigman, Y., Yang, M., Ranzato, M., 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)
60. Tan, X., Triggs, B.: Enhanced local texture feature sets for face recognition under difficult lighting conditions. IEEE transactions on image processing **19**(6), 1635–1650 (2010)
61. Tran, L., Yin, X., Liu, X.: Disentangled representation learning gan for pose-invariant face recognition. In: IEEE Conference on Computer Vision and Pattern Recognition, vol. 3, p. 7 (2017)
62. Turk, M., Pentland, A.: Eigenfaces for recognition. Journal of cognitive neuroscience **3**(1), 71–86 (1991)
63. Wagner, A., Wright, J., Ganesh, A., Zhou, Z., Mobahi, H., Ma, Y.: Toward a practical face recognition system: robust alignment and illumination by sparse representation. IEEE Transactions on Pattern Analysis and Machine Intelligence **34**(2), 372–386 (2012)
64. Wang, H., Li, S.Z., Wang, Y.: Generalized quotient image. In: Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on, vol. 2, pp. II–II. IEEE (2004)
65. Wang, X., Han, T.X., Yan, S.: An hog-lbp human detector with partial occlusion handling. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 32–39 (2009). DOI 10.1109/ICCV.2009.5459207
66. Wen, Y., Zhang, K., Li, Z., Qiao, Y.: A discriminative feature learning approach for deep face recognition. In: B. Leibe, J. Matas, N. Sebe, M. Welling (eds.) Computer Vision – ECCV 2016, pp. 499–515. Springer International Publishing, Cham (2016)
67. Wiskott, L., Krüger, N., Kuiger, N., Von Der Malsburg, C.: Face recognition by elastic bunch graph matching. IEEE Transactions on pattern analysis and machine intelligence **19**(7), 775–779 (1997)
68. Xie, X., Zheng, W.S., Lai, J., Yuen, P.C., Suen, C.Y.: Normalization of face illumination based on large-and small-scale features. IEEE Transactions on Image Processing **20**(7), 1807–1821 (2011)
69. Yang, J., Frangi, A.F., Yang, J.y., Zhang, D., Jin, Z.: Kpca plus lda: a complete kernel fisher discriminant framework for feature extraction and recognition. IEEE Transactions on pattern analysis and machine intelligence **27**(2), 230–244 (2005)
70. Zhang, Y., Shao, M., Wong, E.K., Fu, Y.: Random faces guided sparse many-to-one encoder for pose-invariant face recognition. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 2416–2423 (2013)
71. Zhao, W., Chellappa, R., Phillips, P.J., Rosenfeld, A.: Face recognition: A literature survey. ACM Computing Surveys **35**(4), 399–458 (2003)
72. Zhou, H., Sadka, A.H.: Combining perceptual features with diffusion distance for face recognition. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews) **41**(5), 577–588 (2011)
73. Zhu, Z., Luo, P., Wang, X., Tang, X.: Deep learning identity-preserving face space. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 113–120 (2013)
74. Zhu, Z., Luo, P., Wang, X., Tang, X.: Multi-view perceptron: a deep model for learning face identity and view representations. Advances in Neural Information Processing Systems (NIPS) pp. 217–225 (2014)