A Comprehensive Survey on Pose-Invariant Face Recognition

Changxing Ding   Dacheng Tao

Centre for Quantum Computation and Intelligent Systems

Faculty of Engineering and Information Technology

University of Technology, Sydney

81-115 Broadway, Ultimo, NSW

Australia

Emails: chx.ding@gmail.com, dacheng.tao@uts.edu.au

15 March 2016

Abstract

The capacity to recognize faces under varied poses is a fundamental human ability that presents a unique challenge for computer vision systems. Compared to frontal face recognition, which has been intensively studied and has gradually matured in the past few decades, pose-invariant face recognition (PIFR) remains a largely unsolved problem. However, PIFR is crucial to realizing the full potential of face recognition for real-world applications, since face recognition is intrinsically a passive biometric technology for recognizing uncooperative subjects. In this paper, we discuss the inherent difficulties in PIFR and present a comprehensive review of established techniques. Existing PIFR methods can be grouped into four categories, i.e., pose-robust feature extraction approaches, multi-view subspace learning approaches, face synthesis approaches, and hybrid approaches. The motivations, strategies, pros/cons, and performance of representative approaches are described and compared. Moreover, promising directions for future research are discussed.

Keywords: Pose-invariant face recognition, pose-robust feature, multi-view learning, face synthesis, survey

1   Introduction

Face recognition has been one of the most intensively studied topics in computer vision for more than four decades. Compared with other popular biometrics such as fingerprint, iris, and retina recognition, face recognition has the potential to recognize uncooperative subjects in a non-intrusive manner. Therefore, it can be applied to surveillance security, border control, forensics, digital entertainment, etc. Indeed, numerous works in face recognition have been completed and great progress has been achieved, from successfully identifying criminal suspects from surveillance cameras[1] to approaching human level performance on the popular Labeled Face in the Wild (LFW) database Taigman et al. (2014); Huang et al. (2007). These successful cases, however, may be unrealistically optimistic as they are limited to near-frontal face recognition (NFFR). Recent studies Li et al. (2014); Zhu et al. (2014a) reveal that the best NFFR algorithms Chen et al. (2013); Taigman et al. (2009); Simonyan et al. (2013); Li et al. (2013) on LFW perform poorly in recognizing faces with large poses. In fact, the key ability of pose-invariant face recognition (PIFR) desired by real-world applications remains largely unsolved, as argued in a recent work Abiantun et al. (2014).

---

[1] http://ilinnews.com/armed-robber-identified-by-facial-recognition-technology-gets-22-years/
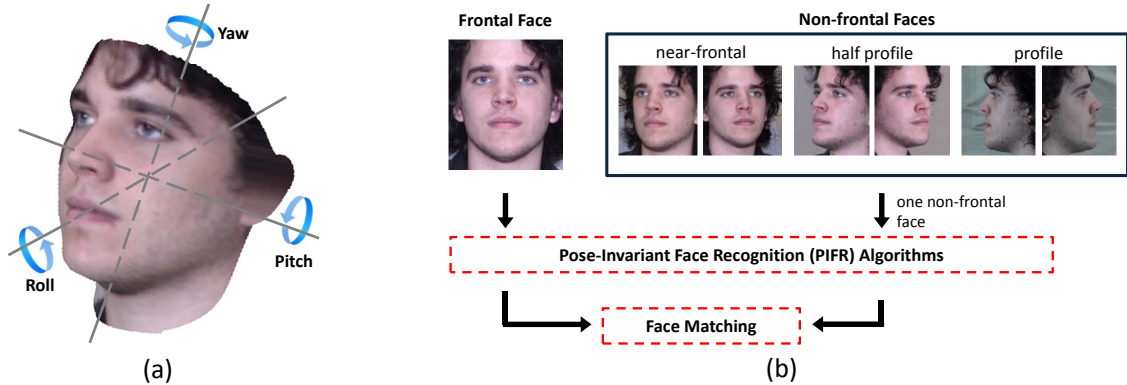
Figure 1: (a) The three degrees of freedom of pose variation of the face, i.e., yaw, pitch, and roll. (b) A typical framework of PIFR. Different from NFFR, PIFR aims to recognize faces captured under arbitrary poses.

PIFR refers to the problem of identifying or authorizing individuals with face images captured under arbitrary poses, as shown in Fig. 1. It is attracting more and more attentions, since face recognition is intrinsically a passive biometric technology for recognizing uncooperative subjects and it is crucial to realize the full potential of face recognition technology for real-world applications. For example, PIFR is important for biometric security control systems in airports, railway stations, banks, and other public places where live surveillance cameras are employed to identify wanted individuals. In these scenarios, the attention of the subjects is rarely focused on surveillance cameras and there is a high probability that their face images will exhibit large pose variations.

The first explorations for PIFR date back to the early 1990s Brunelli and Poggio (1993); Pentland et al. (1994); Beymer (1994). Nevertheless, the substantial facial appearance change caused by pose variation continues to challenge the state-of-the-art face recognition systems. Essentially, it results from the complex 3D structure of the human head. In detail, it presents the following challenges as illustrated in Fig. 2:

- The rigid rotation of the head results in self-occlusion, which means there is loss of information for recognition.

- The position of facial texture varies nonlinearly following the pose change, which indicates the loss of semantic correspondence in 2D images.

- The shape of facial texture is warped nonlinearly along with the pose change, which causes serious confusion with the inter-personal texture difference.

- The pose variation is usually combined with other factors to simultaneously affect face appearance. For example, subjects being captured at a long distance tend to exhibit large pose variations, as they are unaware of the cameras. Therefore, low resolution as well as illumination variations occurs together with large pose variations.

For these reasons, the appearance change caused by pose variation often significantly surpasses the intrinsic differences between individuals. In consequence, it is not possible or effective to directly compare two images under different poses, as in conventional face recognition algorithms. Explicit strategies are required to bridge the cross-pose gap. In recent years, a wide variety of approaches have been proposed which can be broadly grouped into the following four categories, handling PIFR from distinct perspectives:

- Those that extract pose-robust features as face representations, so that conventional classifiers can be employed for face matching.

- Those that project features of different poses into a shared latent subspace where the matching of the faces is meaningful.
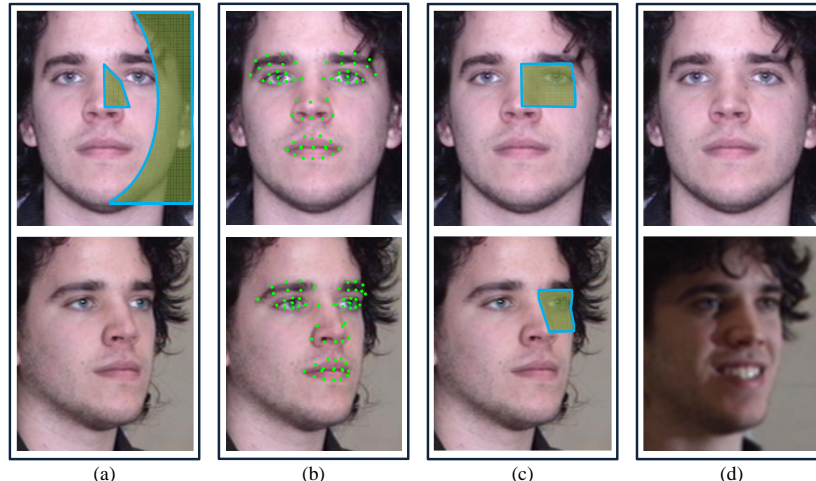
Figure 2: The challenges for face recognition caused by pose variation. (a) self-occlusion: the marked area in the frontal face is invisible in the non-frontal face; (b) loss of semantic correspondence: the position of facial textures varies nonlinearly following the pose change; (c) nonlinear warping of facial textures; (d) accompanied variations in resolution, illumination, and expression.

- Those that synthesize face images from one pose to another pose, so that two faces originally in different poses can be matched in the same pose with traditional frontal face recognition algorithms.

- Those that combine two or three of the above techniques for more effective PIFR.

The four categories of approaches will be discussed in detail in later sections. Inspired by Ouyang et al. (2014), we unify the four categories of PIFR approaches in the following formulation.

$$M\left[W^a F\left(S^a(\mathbf{I}_i^a)\right), W^b F\left(S^b(\mathbf{I}_j^b)\right)\right],\tag{1}$$

where $\mathbf{I}_i^a$ and $\mathbf{I}_j^b$ stand for two face images in pose a and pose b, respectively; S^a and S^b are synthesis operations, after which the two face images are under the same pose; F denotes pose-robust feature extraction; W^a and W^b correspond to feature transformations learnt by multi-view subspace learning algorithms; and M means a face matching algorithm, e.g., the nearest neighbor (NN) classifier. It is easy to see that the first three categories of approach focus their effort on only one operation in Eq. 1. For example, the multi-view subspace learning approaches provide strategies for determining the mappings W^a and W^b; the face synthesis-based methods are devoted to solving S^a and S^b. The hybrid approaches may contribute to two or more steps in Eq. 1. Table 1 provides a list of representative approaches for each of the categories.

The remainder of the paper is organized as follows: Section 2 briefly reviews related surveys for face recognition. Methods based on pose-robust feature extraction are described and analyzed in Section 3. The multi-view subspace learning approaches are reviewed in Section 4. Face synthesis approaches based on 2D and 3D techniques are respectively illustrated in Section 5 and Section 6. The description of hybrid approaches follows in Section 7. The performance of the reviewed approaches is evaluated in Section 8. We close this paper in Section 9 by drawing some overall conclusions and making recommendations for future research.

2   Related Works

Numerous face recognition methods have been proposed due to the non-intrusive advantage of face recognition as a biometric technique. Several surveys have been published. To name a few, good surveys exist for illumination-invariant face recognition Zou et al. (2007), 3D face recognition Bowyer et al. (2006), single image-based face recognition Tan et al. (2006), video-based face recognition Barr et al. (2012), and heterogeneous face recognition Ouyang et al. (2014). There are also comprehensive surveys on various aspects of

face recognition Zhao et al. (2003). Of the existing works, the survey on face recognition across pose Zhang and Gao (2009) that summarizes PIFR approaches before 2009 is the most relevant to this paper. However, there are at least two reasons why a new survey on PIFR is imperative.

First, PIFR has become a particularly important and urgent topic in recent years as the attention of the face recognition filed shifts from research into NFFR to PIFR. The growing importance of PIFR has stimulated a more rapid developmental cycle for novel approaches and resources. The increased number of PIFR publications over the last few years suggests new insights for PIFR, making a new survey for these methods necessary.

Second, the large-scale datasets for PIFR, i.e., Multi-PIE Gross et al. (2010) and IJB-A Klare et al. (2015), have only been established and made available in recent years, creating the possibility of evaluating the performance of existing approaches in a relatively accurate manner. In comparison, approaches reviewed in Zhang and Gao (2009) lack comprehensive evaluation as they use only small databases, where the performance of many approaches have saturated.

This survey spans about 130 most innovative papers on PIFR, with more than 75% of them published in the past seven years. This paper categorizes these approaches from more systematic and comprehensive perspectives compared to Zhang and Gao (2009), reports their performance on newly-developed large-scale datasets, explicitly analyzes the relative pros and cons of different categories of methods, and recommends future development suggestions. Besides, PIFR approaches that require more than two face images per subject for enrollment are not included in this paper, as single image-based face recognition Tan et al. (2006) dominates the research in the past decade. Instead, we direct readers to Zhang and Gao (2009) the for good review on representative works Georghiades et al. (2001); Levine and Yu (2006); Singh et al. (2007).

3 Pose-robust feature extraction

If the extracted features are pose-robust, then the difficulty of PIFR will be relieved. Approaches in this category focus on designing a face representation that is intrinsically robust to pose variation while remaining discriminative to the identity of subjects. According to whether the features are extracted by manually designed descriptors, or by trained machine learning models, the approaches reviewed in this section can be grouped into engineered features and learning-based features.

3.1 Engineered Features

Algorithms designed for frontal face recognition Turk and Pentland (1991); Ahonen et al. (2006) assume tight semantic correspondence between face images, and they directly extract features from the rectangular region of a face image. However, as shown in Fig. 2(b), one of the major challenges for PIFR is the loss of semantic correspondence in the face images. To handle this problem, the engineered features reviewed in this subsection explicitly re-establish the semantic correspondence in the process of feature extraction, as illustrated in Fig 3. Depending on whether facial landmark detection is required, approaches reviewed in this subsection are further divided into landmark detection-based methods and landmark detection-free methods.

3.1.1 Landmark Detection-based Methods

Early PIFR approaches Brunelli and Poggio (1993); Pentland et al. (1994) realized semantic correspondence across pose at the facial component-level. In Pentland et al. (1994), four sparse landmarks, i.e., both eye centers, the nose tip, and the mouth center, are first automatically detected. Image regions containing the facial components, i.e., eyes, nose, and mouth, are estimated and the respective features are extracted. The set of facial component-level features compose the pose-robust face representation. Works that adopt similar ideas include Cao et al. (2010); Zhu et al. (2014b).

Better semantic correspondence across pose is achieved at the landmark-level. Wiskott et al. (1997) proposed the Elastic Bunch Graph Matching (EBGM) model which iteratively deforms to detect dense landmarks. Gabor magnitude coefficients at each landmark are extracted as the pose-robust feature. Similarly, Biswas et al. (2013) described each landmark with SIFT features Lowe (2004) and concatenated the SIFT features of all landmarks as the face representation. More recent engineered features benefit from the
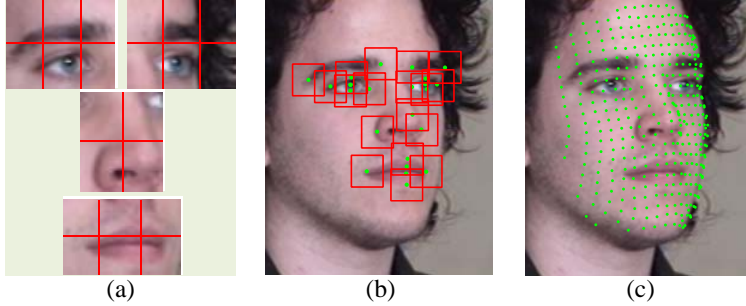
Figure 3: Feature extraction from semantically corresponding patches or landmarks. (a) Semantic correspondence realized in facial component-level Brunelli and Poggio (1993); Pentland et al. (1994); (b) Semantic correspondence by detecting dense facial landmarks Wiskott et al. (1997); Chen et al. (2013); Ding et al. (2015); (c) Tight semantic correspondence realized with various techniques, e.g., 3D face model Li et al. (2009); Yi et al. (2013), GMM Li et al. (2013), MRF Arashloo and Kittler (2011), and stereo matching Castillo and Jacobs (2011b).

rapid progress in facial landmark detection Wang et al. (2014b), which makes dense landmark detection more reliable. For example, Chen et al. (2013) extracted multi-scale Local Binary Patterns (LBP) features from patches around 27 landmarks. LBP features for all patches are concatenated to become a high-dimensional feature vector as the pose-robust feature. A similar idea is adopted for feature extraction in Prince et al. (2008); Zhang et al. (2013a).

Intuitively, the larger the number of landmarks employed, the tighter semantic correspondence that can be achieved. Li et al. (2009) proposed the detection of a number of landmarks with the help of a generic 3D face model. In comparison, Yi et al. (2013) proposed a more accurate approach by employing a deformable 3D face model with 352 pre-labeled landmarks. Similar to Li et al. (2009), the 2D face image is aligned to the deformable 3D face model using the weak perspective projection model, after which the dense landmarks on the 3D model are projected to the 2D image. Lastly, Gabor magnitude coefficients at all landmarks are extracted and concatenated as the pose-robust feature.

Concatenating the features of all landmarks across the face brings about highly non-linear intra-personal variation. To relieve this problem, Ding et al. (2015) combined the component-level and landmark-level methods. In their approach, the Dual-Cross Patterns (DCP) Ding et al. (2015) features of landmarks belonging to the same facial component are concatenated as the description of the component. The pose-robust face representation incorporates a set of features of facial components.

While the above methods crop patches centered around facial landmarks, Fischer et al. (2012) found that the location of the patches for non-frontal faces has a noticeable impact on the recognition results. For example, the positions of patches around some landmarks, e.g., the nose tip and mouth corners, for face images of extreme pose should be adjusted so that fewer background pixels are included.

3.1.2 Landmark Detection-free Methods

The accuracy and reliability of dense landmark detection are critical for building semantic correspondence. However, accurate landmark detection in unconstrained images is still challenging. To handle this problem, (Zhao and Gao 2009; Liao et al. 2013a; Weng et al. 2013; Li et al. 2015) proposed landmark detection-free approaches to extract features around the so-called facial keypoints. For example, Liao et al. (2013a) proposed the extraction of Multi-Keypoint Descriptors (MKD) around keypoints detected by SIFT-like detectors. The correspondence between keypoints among images is established via sparse representation-based classification (SRC). However, the dictionary of SRC in this approach is very large, resulting in an efficiency problem in practical applications. In comparison, Weng et al. (2013) proposed the Metric Learned Extended Robust Point set Matching (MLERPM) approach to efficiently establish the correspondence between the keypoints of two faces.

Similarly, Arashloo et al. (2011) proposed an landmark detection-free approach based on Markov Random Field (MRF) to match semantically corresponding patches between two images. In their approach, the

densely sampled image patches are represented as the nodes of the MRF model, while the 2D displacement vectors are treated as labels. The goal of the MRF-based optimization is to find the assignment of labels with minimum cost, taking both translations and projective distortions into consideration. The matching cost between patches can be measured from the gradient Arashloo and Kittler (2011), or the gradient-based descriptors Arashloo and Kittler (2013). The main shortcoming of this approach lies in the high computational burden in the optimization procedure, which is accelerated by GPUs in their other works Arashloo and Kittler (2013); Rahimzadeh Arashloo and Kittler (2014). For face recognition, local descriptors are employed to extract features from semantically corresponding patches Arashloo and Kittler (2013, 2014).

Another landmark detection-free approach is the Probabilistic Elastic Matching (PEM) model proposed by Li et al. (2013). Briefly, PEM first learns a Gaussian Mixture Model (GMM) from the spatial-appearance features Wright and Hua (2009) of densely sampled image patches in the training set. Each Gaussian component stands for patches of the same semantic meaning. A testing image is represented as a bag of spatial-appearance features. The patch whose feature induces the highest probability on each Gaussian component is found. Concatenating the feature vectors of these patches forms the representation of the face. Since all testing images follow the same procedure, the semantic correspondence is established. They reported improved performance on LFW by establishing the semantic correspondence. However, PEM has the disadvantage in efficiency, since the semantic correspondence is inferred through GMM. As GMM only plays the role of a bridge to establish semantic correspondence between two images and the extracted features are still engineered, we classify PEM as an engineered pose-robust feature.

To achieve pixel-wise correspondence between two face images under different poses, Castillo and Jacobs (2007, 2009, 2011b) explored the stereo matching algorithm. In their approach, four facial landmarks are first utilized to estimate the epipolar geometry of the two faces. The correspondence between pixels of the two face images is then solved by a dynamic programming-based stereo matching algorithm. Once the correspondence is known, normalized correlation based on raw image pixels is used to calculate the similarity score for each pair of corresponding pixels. The summation of the similarity scores of all corresponding pixel pairs forms the similarity score of the image pair. In another of their works Castillo and Jacobs (2011a), they replace raw image pixels with image descriptors to calculate the similarity of pixel pairs and fuse the similarity scores using Support Vector Machine (SVM).

The engineered features handle the PIFR problem only from the perspective of establishing semantic correspondences, which has clear limitations. First, semantic correspondence may be completely lost due to self-occlusion in large pose images. To cope with this problem, Arashloo and Kittler (2011) and Yi et al. (2013) proposed the extraction of features only from the less-occluded half faces. Second, the engineered features cannot relieve the challenge caused by the nonlinear warping of facial textures due to pose variation. Therefore, the engineered features can generally handle only moderate pose variations.

3.2 Learning-based Features

The learning-based features are extracted by machine learning models that are usually pre-trained by multi-pose training data. Compared with the engineered features, the learning-based features are better at handling the problem of self-occlusion and non-linear texture warping caused by pose variations.

Inspired by their impressive ability to learn high quality image representations, neural networks have recently been employed to extract pose-robust features, as illustrated in Fig. 4. Zhu et al. (2013) designed a deep neural network (DNN) to learn the so called Face Identity-Preserving (FIP) features. The deep network is the stack of two main modules: the feature extraction module and the frontal face reconstruction module. The former module has three locally connected convolution layers and two pooling layers stacked alternately. The latter module contains a fully-connected reconstruction layer. The input of the model is a set of pose-varied images of an individual. The output of the feature extraction module is employed to recover frontal faces through the latter module, therefore the frontal face is saved as a supervised signal to train the network. The logic of this method is that regardless of the pose of the input image, the output of the reconstruction module is encouraged to be as close as possible to the frontal pose image of the subject. Thus, the output of the feature extraction module must be pose-robust. Due to the deep structure of the model, the network has millions of parameters to tune and therefore requires a large amount of multi-pose training data.

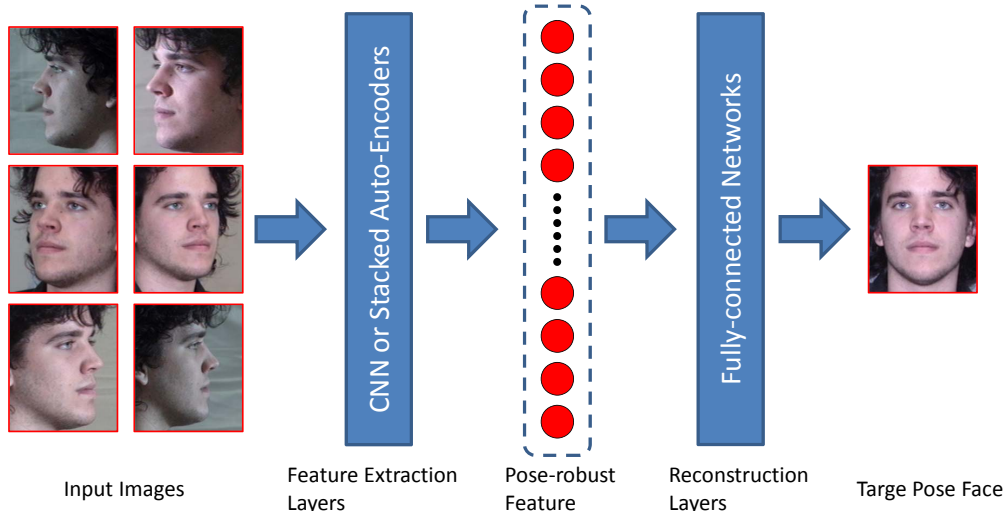Another contemporary work Zhang et al. (2013a) adopted a single-hidden-layer auto-encoder to extract

Figure 4: The common framework of deep neural network-based pose-robust feature extraction methods Zhu et al. (2013); Zhang et al. (2013a); Kan et al. (2014).

pose-robust features. Compared with Zhu et al. (2013), the neural network built in Zhang et al. (2013a) is shallow because it contains only one single-hidden layer. Like Zhu et al. (2013), the input of the network is a set of pose-varied images of an individual, but the target signal of the output layer is more flexible than Zhu et al. (2013), i.e., it could be the frontal pose image of the identity or a random signal that is unique to the identity. As argued by the authors, the target value for the output layer is essentially an identity representation which is not necessarily a frontal face, therefore the vector that is represented by the neurons in the hidden layer can be used as the pose-robust feature. Due to the shallow structure of the network, less amount of training data is required for training compared with Zhu et al. (2013). However, it may not extract as high-quality pose-robust feature as Zhu et al. (2013). To handle this problem, the authors proposed building multiple networks of exactly the same structure. The input of the networks is the same, while their target values are different random signals. In this way, parameters of the learnt network models are different, and the multiple networks randomly encode a variety of information about the identity. The final pose-robust feature is the vertical pile of the hidden layer outputs of all networks.

Kan et al. (2014) proposed the stacked progressive auto-encoders (SPAE) model to learn pose-robust features. In contrast to Zhang et al. (2013a), SPAE stacks multiple shallow auto-encoders, thus it is a deep network. The authors argue that the direct transformation from the non-frontal face to the frontal face is a complex non-linear transform, thus the objective may be trapped into local minima because of its large search region. Inspired by the observations that pose variations change non-linearly but smoothly, the authors proposed learning pose-robust features by progressive transformation from the non-frontal face to the frontal face through the stack of several shallow auto-encoders. The function of each auto-encoder is to map the face images in large poses to a virtual view in slighter pose changes, and meanwhile keep those images already in smaller poses unchanged. In this way, the deep network is forced to approximate its eventual goal by several successive and tractable tasks. Similar to Zhu et al. (2013), the output of the top hidden layers of SPAE is used as the pose-robust feature.

The above three networks are single-task based, i.e., the extracted pose-robust features are required to reconstruct the face image under a single target pose. In comparison, Yim et al. (2015) designed a series interconnection network which includes a main DNN and an auxiliary DNN. The pose-robust feature extracted by the main DNN is required not only to reconstruct the face image under the target pose, but also recover the original input face image with the auxiliary DNN. With the multi-task strategy, the identity-preserving ability of the extracted pose-robust features is observed to be enhanced compared with single-task based DNN.

Apart from the deep neural networks, a number of other machine learning models are utilized to extract pose-robust features. For example, kernel-based models, e.g., Kernel PCA Liu (2004) and Kernel LDA Kim

and Kittler (2006); Huang et al. (2007); Tao et al. (2006, 2007, 2009), were employed to learn nonlinear transformation to a high-dimensional feature space where faces of different subjects are assumed to be linearly separable, despite of pose variation. However, this assumption may not necessarily hold in real applications Zhang and Gao (2009). Besides, it has been shown that the coefficients for some face synthesis models Chai et al. (2007); Annan et al. (2012); Blanz and Vetter (2003) which will be reviewed in Sections 5 and 6 can be regarded as pose-robust features for recognition. Their common shortcoming is that they suffer from statistical stability problems in image fitting due to the complex variations that appear in real images.

Another group of learning-based approaches is based on the face-similarity of one face image to N template subjects Müller et al. (2007); Schroff et al. (2011); Liao et al. (2013b). Each of the template subjects has a number of pose-varied face images. The pose-robust representation of the input image is also N dimensional. In Liao et al. (2013b), the kth element of the representation measures the similarity of the input image to the kth template subject. This similarity score is obtained by first computing the convolution of the input image's low-level features with those of the kth template subject, and then pooling the convolution results. It is expected that the pooling operation will lead to robustness to the nonlinear pose variations. In comparison, Schroff et al. (2011) proposed the Doppelgänger list approach to sort the template subjects according to their similarity to the input face image. The sorted Doppelgänger list is utilized as the pose-robust feature of the input image. Besides, Kafai et al. (2014) proposed the Reference Face Graph (RFG) approach to measure the discriminative power of each of the template subjects. The similarity score of the input image to each template subject is modified by weighting the discriminative power of the template subject. Their experiments demonstrate that performance is improved using this weighting strategy. Compared with Zhu et al. (2013); Zhang et al. (2013a); Kan et al. (2014), the main advantage of the two approaches Liao et al. (2013b); Kafai et al. (2014) is that they have no free parameters.

3.3 Discussion

The engineered features achieve pose robustness by re-establishing the semantic correspondence between two images. The semantic correspondence cannot handle the challenge of self-occlusion or nonlinear facial texture warping caused by pose variation. The learning-based features compensate for this shortcoming by utilizing non-linear machine learning models, e.g., deep neural networks. These machine learning models may produce higher quality pose-robust features, but this is usually at the cost of massive labeled multi-pose training data, which is not necessarily available in practical applications Liu and Tao (2015). The capacity of the learning-based features may be further enhanced by combining the benefit of semantic correspondence, e.g., extracting features from semantically corresponding patches rather than the holistic face image.

4   Multi-view Subspace Learning

The pose-varied face images are distributed on a highly nonlinear manifold Tenenbaum et al. (2000), which greatly degrades the performance of traditional face recognition models that are based on the single linear subspace assumption Turk and Pentland (1991). The multi-view subspace learning-based approaches reviewed in this section tackle this problem by dividing the nonlinear manifold into a discrete set of pose spaces and regard each pose as a single view, and pose-specific projections to a latent subspace shared by different poses are subsequently learnt (Kim et al. 2003; Prince and Elder 2005). Since the images of one subject are captured under different poses of the same face, they should be highly correlated in this subspace; therefore face matching can be performed due to feature correspondence. According to the properties of the models used, the approaches reviewed in this section are divided into linear models and nonlinear models. An illustration of the multi-view subspace learning framework is shown in Fig 5.

4.1   Linear Models

4.1.1   Discriminative Linear Models

Li et al. (2009) proposed learning the multi-view subspace by exploiting Canonical Correlation Analysis (CCA). The principle of CCA is to learn two projection matrices, one for each pose, to project the samples
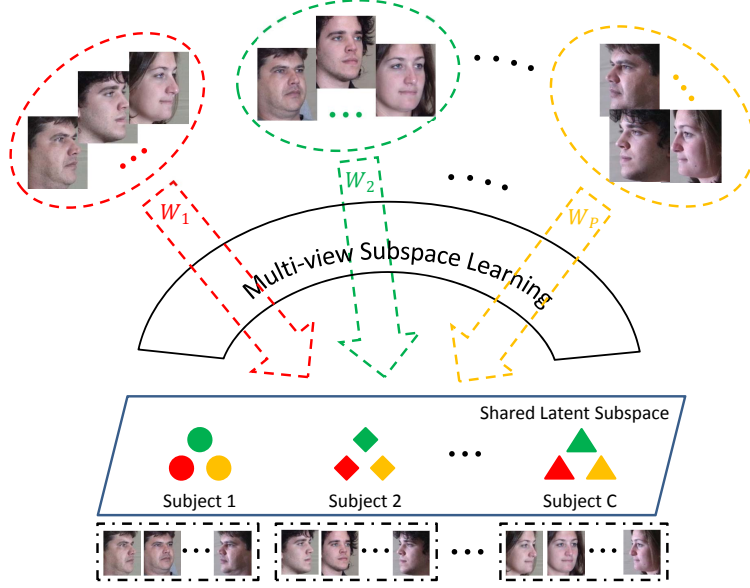
Figure 5: The framework of multi-view subspace learning-based PIFR approaches Kan et al. (2012); Prince et al. (2008); Li et al. (2009); Sharma et al. (2012b). The continuous pose range is divided into P discrete pose spaces, and pose-specific projections, i.e., $W_1, W_2, \cdots, W_P$, to the latent subspace are learnt. Approaches reviewed in this section differ in the optimization of the projections.

of the two poses into a common subspace, where the correlation between the projected samples from the same subject is maximized. Formally, given N pairs of samples from two poses: $\{(x_{11}, x_{21}), (x_{12}, x_{22}), \cdots, (x_{1N}, x_{2N})\}$, where $x_{pi} \in R^{d_p}$ represents the data of the pth pose from the ith pair, and $1 \leq p \leq 2$, $1 \leq i \leq N$. It is required that the two samples in each pair belong to the same subject. Two matrices $X_1 = [x_{11}, x_{12}, \cdots, x_{1N}]$ and $X_2 = [x_{21}, x_{22}, \cdots, x_{2N}]$ are defined to represent the data from the two poses, respectively. Two linear projections w_1 and w_2 are learnt for X_1 and X_2, respectively, such that the correlation of the low-dimensional embeddings $w_1^T X_1$ and $w_2^T X_2$ is maximized:

$$\max_{w_1, w_2} corr[w_1^T X_1, w_2^T X_2]$$
$$s.t. \ \|w_1\| = 1, \|w_2\| = 1. \tag{2}$$

By employing the Lagrange multiplier, the above problem can be solved by the generalized eigenvalue decomposition method. Since the projection of face vectors by CCA leads to feature correspondence in the shared subspace, the subsequent face recognition can be conducted.

Considering the fact that CCA emphasizes only the correlation but ignores data variation in the shared subspace, which may affect face recognition performance, Sharma and Jacobs (2011) and Li et al. (2011) proposed the use of Partial Least Squares (PLS) to learn the multi-view subspace for both poses. Formally, PLS finds the linear projections w_1 and w_2 such that

$$\max_{w_1, w_2} cov[w_1^T X_1, w_2^T X_2]$$
$$s.t. \ \|w_1\| = 1, \|w_2\| = 1. \tag{3}$$

Recall that the relation between correlation and covariance is as follows,

$$corr[w_1^T X_1, w_2^T X_2] = \frac{cov[w_1^T X_1, w_2^T X_2]}{std(w_1^T X_1) std(w_2^T X_2)}, \tag{4}$$

where $std(\cdot)$ stands for standard deviation. It is clear that PLS tries to correlate the samples of the same subject as well as capture the variations present in the original data, which helps to enhance the ability to

differentiate the training samples of different subjects in the shared subspace Sharma and Jacobs (2011). Therefore, better performance by PLS than CCA was reported in Sharma and Jacobs (2011).

In contrast to CCA and PLS, which can only work in the scenario of two poses, Rupnik and Shawe-Taylor (2010) presented the Multi-view CCA (MCCA) approach to obtain one common subspace for all P poses available in the training set. In MCCA, a set of projection matrices, one for each of the P poses, is learnt based on the objective of maximizing the sum of all pose pair-wise correlations:

$$\max_{w_1,\cdots,w_P} \sum_{i<j} corr[w_i^T X_i, w_j^T X_j]$$
$$s.t. \ \|w_i\| = 1, i = 1, \cdots, P \tag{5}$$

The aforementioned methods only concern pair-wise closeness in the shared subspace and do not make use of the label information. In contrast to the above methods, Sharma et al. (2012a) presented a two-stage framework for multi-view subspace learning. In this method, the first stage learns pose-specific linear projections using MCCA, after which all training samples are projected to the common subspace, which is assumed to be linear. The second stage learns discriminative projections in the shared subspace using Linear Discriminant Analysis (LDA). The experimental results indicate that the two-stage approach consistently results in improvements to performance, compared to the unsupervised methods.

Sharma et al. (2012b) further proposed a general framework named Generalized Multiview Analysis (GMA) for multi-view subspace learning. The contribution of GMA is two-fold. First, existing methods based on generalized eigenvalue decomposition, e.g., PCA, CCA, PLS, and MCCA, can be unified in this framework. Second, existing single-view models can be extended to their multi-view versions under this framework. For example, the LDA model can be extended to its multi-view counterpart, Generalized Multi-view LDA (GMLDA). Technically, GMLDA makes a tradeoff between the discriminability within each pose and the correlation between poses:

$$\max_{w_1,\cdots,w_P} \sum_{i=1}^{P} \mu_i w_i^T S_{bi} w_i + \sum_{i<j} \lambda_{ij} w_i^T Z_i Z_j^T w_j$$
$$s.t. \ \sum_{i=1}^{P} \gamma_i w_i^T S_{wi} w_i = 1, \tag{6}$$

where μ_i, λ_{ij}, and γ_i are model parameters. S_{bi} and S_{wi} are the between-class scatter matrix and the within-class matrix, respectively, for the ith pose. Therefore, the first term in the objective function enhances the discriminative power within each pose. Z_i are defined as the matrix with columns that are class means for the ith pose, and corresponding columns of Z_i and Z_j should be of the same subject. In this way, inter-pose faces of the same subject are correlated and clustered in the shared subspace, and the gap caused by pose variation is thus reduced. The experimental results in Sharma et al. (2012b) reveal that GMLDA outperforms their previously proposed two-stage method Sharma et al. (2012a). Works that adopt a similar idea to GMA includes Huang et al. (2013); Guo et al. (2014) which formulate multi-view subspace learning methods in graph embedding frameworks.

Methods based on CCA or PLS usually require each training subject to have the same number of faces for all poses, a condition which may not be satisfied in real applications. GMLDA relieves this demanding requirement but it still requires that each pose pair (Z_i, Z_j) in Eq. 6 has exactly the same training subjects. To handle this problem, Kan et al. (2012) proposed the Multi-view Discriminant Analysis (MvDA) approach which utilizes all face images from all poses. In contrast to GMLDA, MvDA builds a single between-class scatter matrix S_b and a single within-class scatter matrix S_w from both the inter-pose and intra-pose embeddings in the shared subspace:

$$S_w = \sum_{i=1}^{C} \sum_{j=1}^{P} \sum_{k=1}^{n_{ij}} (y_{ij}^k - \mu_i)(y_{ij}^k - \mu_i)^T,$$
$$S_b = \sum_{i=1}^{C} n_i (\mu_i - \mu)(\mu_i - \mu)^T, \tag{7}$$

where y_{ij}^k stands for low-dimensional embedding of the kth sample from the jth pose of the ith subject. μ_i is the mean of the low-dimensional embeddings of the ith subject, and μ is the mean of the low-dimensional

embeddings of all C subjects. The objective of MvDA is similar to that of LDA, i.e., to maximize the ratio between S_b and S_w. The pose-specific projections are obtained by formulating the objective as the optimization of a generalized Rayleigh quotient. As variations from both inter-pose and intra-pose faces are considered together in the objective function, the authors argued that a more discriminative subspace can be learnt.

4.1.2 Generative Linear Models

In addition to the above discriminative models, generative models are also explored for multi-view subspace learning. One typical model is the Tied Factor Analysis (TFA) approach proposed by Prince et al. (2008). The core of TFA is the assumption that there exists an idealized identity subspace where each multi-dimensional variable h_i represents the identity of one subject, regardless of the pose. Suppose x_{ij}^k stands for the kth sample from the jth pose of the ith subject, then x_{ij}^k is generated by the pose-specific linear transformations of h_i by W_j in addition to the offset μ_j and Gaussian noise $\epsilon_{ij}^k \sim \mathcal{G}(0, \Sigma_j)$:

$$x_{ij}^k = W_j h_i + \mu_j + \epsilon_{ij}^k. \tag{8}$$

The model parameters W_j, μ_j, and Σ_j $(1 \leq j \leq P)$ are estimated from multi-pose training data by using the Expectation-Maximization (EM) algorithm. In the recognition step, TFA calculates the probability that two images under different poses are generated from the same identity vector h_i under the linear transformation scheme.

Cai et al. (2013) proposed the Regularized Latent Least Square Regression (RLLSR) method, which is similar to Prince's work Prince et al. (2008). RLLSR is based on the same assumption as TFA, however, it reformulates Eq. 8 in a least square regression framework, where the observed images x_{ij}^k is treated as regressor and the identity variable h_i as response. The identity variables h_i and the model parameters W_j and μ_j are estimated by the RLLSR model with an alternating optimization scheme. To overcome overfitting and enhance the generalization ability, proper regularizations are imposed on the objective function of RLLSR. In the recognition phase, the cosine metric is adopted to measure the similarity between the estimated identity variables.

The assumption that a single identity variable h_i can be used to faithfully generate all different images of a subject under a certain pose seems to be over-simplified. It is well-known that the appearance of a subject changes substantially as a result of the rich variations in illumination and expression and so forth, even if the subject stays in the same pose. In other words, TFA may not be able to effectively handle the complex within-class variations that appear in each pose space. To overcome this limitation, Li et al. (2012) proposed the Tied Probabilistic Linear Discriminant Analysis (Tied-PLDA) approach. Tied-PLDA is built on TFA but incorporates both within-class and between-class variations in the model:

$$x_{ij}^k = W_j h_i + G_j w_{ij}^k + \mu_j + \epsilon_{ij}^k, \tag{9}$$

where the respective definitions of x_{ij}^k, h_i, μ_j, and ϵ_{ij}^k are the same as those in TFA. W_j and G_j are pose-specific transformations which account for between-class subspace and within-class subspace, respectively. All images belonging to the ith subject share the same identity variable h_i, but each image has a different variable w_{ij}^k which represents the position in the within-class subspace. As in TFA, Tied-PLDA optimizes the model parameters using the EM algorithm. In the recognition step, it calculates the probability that two images are generated from the same identity vector h_i, regardless of whether they are in the same pose. As the face generation process is formulated in a more reasonable way by Tied-PLDA, better performance than TFA was reported Li et al. (2012).

4.2 Nonlinear Models

Appearance changes in face images are highly nonlinear due to the substantial local warp and occlusion caused by pose variation. The representational ability of the linear methods is limited, thus these methods may not be able to convert data from different poses into an ideal common space. From this perspective, nonlinear techniques are preferable.
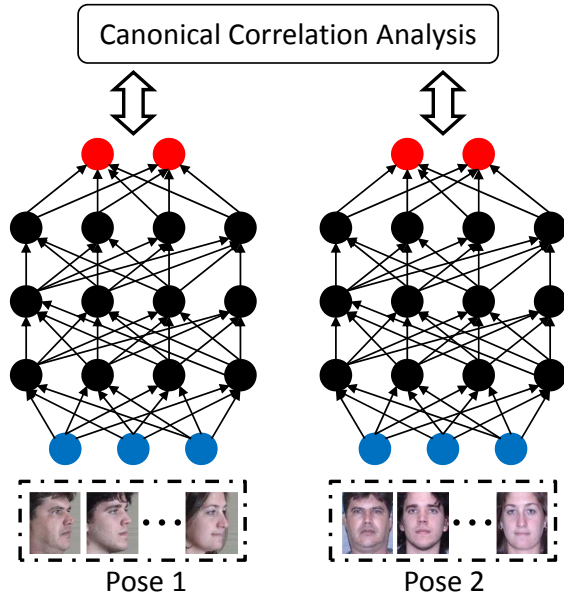
Figure 6: The scheme of the Deep Canonical Correlation Analysis (DCCA) method Andrew et al. (2013). One network is learnt for each pose (view) and the outputs of the two networks are maximally correlated.

A natural nonlinear extension of the introduced linear models is realized via the kernel technique, such that the nonlinear classification problem in the original space is converted to be the linear classification problem in the higher dimensional kernel space. For example, Akaho (2006) proposed the extension of CCA to Kernel CCA (KCCA). Sharma et al. (2012b) extended the GMLDA approach to Kernel GMLDA (KGMLDA). Better performance of KCCA than CCA on the face data was observed by Wang et al. (2014a).

Inspired by the ability of deep learning models to learn nonlinear representations, a recent topic of interest has been to design multi-view subspace learning methods via deep structures. Andrew et al. (2013) proposed the Deep Canonical Correlation Analysis (DCCA) method, which is another nonlinear extension of CCA. In brief, DCCA builds one deep network for each of the two poses, and the representations of the highest layer of the networks are constrained to be correlated, as illustrated in Fig 6. Compared with KCCA, the training time of DCCA scales well with the size of the training set, and if is not necessary to reference the training data in the testing stage. It was reported in Andrew et al. (2013) that higher performance was achieved by DCCA than either CCA or KCCA.

DCCA is an unsupervised method, which means that it may not be suitable for classification tasks. Wang et al. (2014a) introduced the Deeply Coupled Auto-encoder Networks (DCAN) method to effectively employ the label information of the training data. Similar to DCCA, DCAN also constructs one deep network for each of the two poses. The two networks are discriminatively coupled with each other in every corresponding layer, which enables samples from two poses to be projected to one common discriminative subspace. The whole network is able to represent complex non-linear transformations as the number of layers increases; therefore, the gap between the two poses narrows and the discriminative capacity of the common subspace is enhanced. The authors reported better performance by DCAN than a number of linear methods, e.g., CCA, PLS, and MvDA, and the nonlinear method KCCA, because of the nonlinear learning capacity of the deep networks.

4.3 Discussion

The multi-view subspace learning approaches reviewed in this section attempt to narrow the gap between different poses by projecting their features to a common subspace with pose-specific transformations. Among the existing techniques, linear models have the advantage in efficiency since the low-dimensional embeddings can be computed directly by matrix multiplication. The capacity of the linear models is limited, however, as the appearance variations resulting from pose changes are intrinsically nonlinear. The nonlinear techniques
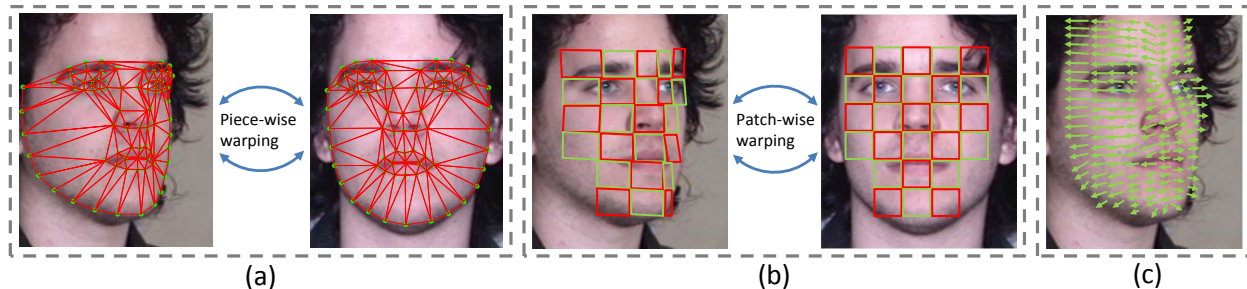
Figure 7: Three main schemes of 2D-based pose normalization methods. (a) Piece-wise warping; (b) Patch-wise warping; (c) Pixel-wise displacement.

make up for this imperfection by learning nonlinear projections, but at the cost of lower efficiency in model training or testing. The deep learning-based nonlinear models also require larger training data. The common shortcoming of the multi-view subspace learning methods is that they depend on large training data which incorporate all the poses that might appear in the testing phase, but the large amount of multi-pose training data might not be available in real-world applications.

5 Face Synthesis based on 2D Methods

Since directly matching two faces under different poses is difficult, one intuitive method is to conduct face synthesis so that the two faces are transformed to the same pose, allowing conventional face recognition algorithms to be used for matching. Existing face synthesis methods for PIFR can be broadly classified into methods based on 2D techniques and methods based on 3D techniques, depending on whether the synthesis is accomplished in the 2D domain or the 3D domain. In this section, we review the 2D techniques which accomplish face synthesis directly in the 2D image domain.

5.1 2D Pose Normalization

5.1.1 Piece-wise Warping

The three main schemes of 2D pose normalization methods are illustrated in Fig. 7. Early 2D pose normalization works for face synthesis are based on the piecewise warping strategy. The piecewise warping approaches transform the shape of the face image in a piecewise manner to another specified pose. Each piece refers to one triangle of the triangular mesh, which is generated by Delaunay triangulation of the dense facial landmarks on the face. The warping transformation between each pair of triangle regions of the original image and the target image can be an affine warping Gao et al. (2009) or a thin-plate splines-based warping Bookstein (1989). The effects of the two warping strategies were compared by Gao et al. (2009).

A number of works directly warp each face image to a uniform shape in the frontal pose with neutral expression Cootes et al. (2001); Gao et al. (2009). González-Jiménez and Alba-Castro (2007) argued that warping all faces to the uniform shape is so strict that the discriminative identity information for recognition may be lost. They therefore collected a training set which incorporated the coordinates of dense facial landmarks. A Point Distribution Model (PDM) Cootes et al. (1995) was then built and the parameters that are assumed to control only pose variations were identified. In the testing phase, the mesh of a testing image can be conveniently transformed to other poses by adjusting its pose parameters in PDM, while the parameters controlling the identity and variations in expression stay the same. In their experiments, the authors observed improved recognition performance over using a single uniform shape for face synthesis. However, the identified pose parameters in PDM may not be well separated, i.e., they may also control the non-rigid transformation by expression or identity, resulting in error in the synthesized face images.

Asthana et al. (2009) proposed the Gaussian Process Regression (GPR) model to learn the correspondence between facial landmarks in the frontal and non-frontal poses offline. In the testing phase, given a frontal face image and its facial landmarks, the landmark coordinates of non-frontal faces to be synthesized are predicted

by GPR, followed by warping the texture of the original frontal image. In this way, the gallery set containing frontal faces is expanded with synthesized non-frontal faces and a probe image can be compared with gallery images in similar pose. However, enlarging the gallery set reduces the efficiency of face recognition systems. In another work Asthana et al. (2011a), the authors trained GPR to predict the landmark coordinates of the virtual frontal pose from the coordinates of a non-frontal face. In this way, all the images are transformed to the frontal pose and compared without expanding the gallery set.

Taigman et al. (2014) proposed the inference of the landmark locations in the virtual frontal pose for each non-frontal face with the help of a generic 3D face model. First, the 3D face model is rotated to the pose of the 2D image by aligning the facial landmarks of both the model and the image. The residuals between the landmarks of the 2D image and the projected locations of the 3D landmarks are then added to the uniform shape in the frontal pose as compensation, which is assumed to reduce the distortion to the identity caused by pose normalization. In spite of the usage of 3D face model, pose normalization still happens in the 2D image domain. Therefore, we categorize this approach as a 2D pose normalization technique.

The above approaches directly infer the personalized shape of a face in a novel pose. Berg and Belhumeur (2012) proposed another method of saving the identity information. For each non-frontal image, triangulation is performed according to the generic landmark locations of all faces in this pose rather than the image's own facial landmarks. The image is then transformed to the uniform shape of the frontal pose by piecewise affine warping. In this way, a smaller distortion of the identity of the test image is achieved. In the experiment on LFW, the authors reported better performance by the 'identity-preserving' warping strategy than the uniform warping method.

Despite its simplicity, the piecewise warping method has obvious limitations. First, the pose of the synthesized face cannot deviate too much from that of the original face because of the risk of causing severe degradation in image quality. Heo and Savvides (2012a) investigated the ability of this method to handle yaw angle variation. They concluded that if the yaw difference exceeds $\pm 15°$, the resulting warped images produce obvious stretching artifacts, caused by over-sampling the texture in relatively small regions. Second, the quality of the synthesized image depends heavily on the accuracy of the detection of each landmark as the warping is determined solely by the landmarks. However, facial landmark detection continues to be a difficult problem for half-profile or profile images.

5.1.2 Patch-wise Warping

In contrast to the above works based on piecewise warping, Ashraf et al. (2008) modeled the face image as a collection of patches and accomplished the reconstruction of the face image using a patch-wise strategy. For each patch, they proposed the "stack-flow" approach for frontal patch synthesis from patches of a non-frontal pose. In the training phase, the "stack-flow" method learns the optimal affine warp using the Lucas-Kanade (LK) algorithm which aligns patches in the stack of the non-frontal faces to corresponding patches in the stack of frontal faces. In the testing phase, each of the patches is warped to the frontal pose with the pre-learnt warp parameters, after which a frontal pose image is synthesized. To stably learn the affine warp between the half-profile patches and the frontal patches, the authors also proposed a composite strategy that successively learns the warp parameters between intermediate poses. Due to the flexibility of the "stack-flow" approach, it is reported that it can cover a wider range of pose variations than piecewise warping approaches.

The patch-wise warps in the "stack-flow" approach are optimized individually for each patch, without considering consistency at the overlapped pixels between two nearby patches. To tackle this problem, Ho and Chellappa (2013) proposed learning a globally optimal set of local warps for frontal face synthesis. This method is composed of two main steps. First, a set of candidate affine warps for each patch is obtained by aligning the patch to corresponding patches of a set of frontal training images, using an improved LK algorithm Ashraf et al. (2010) that is robust to illumination variations. Second, a globally optimal set of local warps $\{p_i\}_{i=1}^{N}$ is chosen from the candidate warps, one for each of the N patches. The problem is formulated into the following optimization problem:

$$\min_{p_1,\cdots,p_N} \sum_{i=1}^{N} E_i(p_i) + \lambda \sum_{(i,j)\in\varepsilon} E_{ij}(p_i, p_j), \tag{10}$$

where $E_i(p_i)$ measures the cost of assigning the affine warp p_i to the ith patch, while $E_{ij}(p_i, p_j)$ is a smoothness term which measures the cost of inconsistency at the region of overlaps between the ith patch and j^{th}

patch. $(i, j) \in \varepsilon$ if the two patches are four-connected neighbors. Therefore, Eq. 10 strikes a balance between the flexibility for each patch and the global consistency for all the patches. Eq. 10 is solved as a discrete labeling problem using MRF Ho and Chellappa (2013).

One common shortcoming for Ashraf et al. (2008) and Ho and Chellappa (2013) is that they divide both frontal and non-frontal images into the same regular set of local patches. This dividing strategy results in the loss of semantic correspondence for some patches when the pose difference is large, as argued in Li et al. (2009), therefore the learnt patch-wise affine warps may lose practical significance. For various methods of detecting patches with semantic correspondence, please refer to Section 3 of this review.

5.1.3 Pixel-wise Displacement

The above piece-based or patch-based pose normalization methods cannot handle the local nonlinear warps that appear in each piece or patch. Beymer and Poggio (1995) proposed the "parallel deformation" approach which predicts the pixel-wise displacement between two poses. This approach first establishes the dense pixel-wise semantic correspondence between images of different poses and subjects using the optical flow-based method. The displacement fields containing dense pixel-wise displacement between two poses from a set of training subjects are recorded as the template displacement fields. Given a testing face image, its displacement field can be estimated by a linear combination of the template displacement fields. With the estimated displacement field, the testing face image is deformed to the image under another pose.

The major drawback of the "parallel deformation" approach lies in the difficulty of establishing pixel-wise correspondence between images. Li et al. (2012, 2014) proposed the generation of the template displacement fields using images synthesized by a set of 3D face models. The pixel-wise correspondence between the synthesized images can be easily inferred via the 3D model vertices, therefore this approach implicitly utilizes 3D facial shape priors for pose normalization. The authors also proposed the implicit Morphable Displacement Field (iMDF) method Li et al. (2012) and the Maximal Likelihood Correspondence Estimation (MLCE) method Li et al. (2014) to effectively estimate the convex combination of the template displacement fields. In the testing phase, the 3D models are discarded and only the template displacement fields are utilized for face synthesis, i.e., the face synthesis is accomplished in the 2D image domain.

5.2 Linear Regression Models

The same as the approaches in Section 4, methods described in this subsection divide the continuous pose space into a set of discrete pose segments. Faces that fall into the same pose segment are assumed to have the same pose p.

The earliest work that formulated face synthesis as a linear regression problem was by Beymer and Poggio (1995). Under the assumption of orthogonal projection and a condition of constant illumination, the holistic face image is represented as the linear combination of a set of training faces in the same pose, and the same combination coefficients are employed for face synthesis under another pose. However, this approach requires dense pixel-wise correspondence between face images, which is challenging in practice. Later works conducted face synthesis in a patch-wise strategy Chai et al. (2007), reducing the difficulty in alignment. Another advantage of the patch-based methods is that each patch can be regarded as a simple planar surface, thus the transformation of the patches across poses can be approximated by linear regression models.

Inspired by Beymer's work Beymer and Poggio (1995), Chai et al. (2007) proposed the Local Linear Regression (LLR) method for face synthesis. LLR works on the patch level, based on the key assumption that the manifold structure of a local patch stays the same across poses. Formally, suppose there exist two training matrices $D_0 = [x_{01}, x_{02}, \cdots, x_{0n}] \in R^{d \times n}$ and $D_p = [x_{p1}, x_{p2}, \cdots, x_{pn}] \in R^{d \times n}$, whose columns are composed of the vectorized local patches in the frontal pose and non-frontal pose p, respectively. Note that the corresponding columns (patches) in D_0 and D_p are of the same subject. In the testing time, given an image patch x_{pt} in pose p, the first step of LLR is to predict x_{pt} from the linear combination of columns in D_p, where the combination coefficients α_t is computed by the least square algorithm:

$$\min_{\alpha_t} \|x_{pt} - D_p \alpha_t\|^2. \tag{11}$$

The second step of LLR is to synthesize the frontal patch x_{0t} with the learnt coefficients α_t:

$$x_{0t} = D_0 \alpha_t. \tag{12}$$

By repeating the above process for all available patches in the testing image, the final virtual frontal pose image can be obtained by overlapping all the predicted frontal patches.

In spite of its simplicity, LLR suffers from the overfitting problem in Eq. 11, which means the manifold structure learnt in pose p may not faithfully represent the structure in another pose. To relieve this problem, several improved approaches have been proposed by imposing regularization terms on Eq. 11. These include the lasso regularization Annan et al. (2012); Zhang et al. (2013b), the ridge regularization Annan et al. (2012), local similarity regularization Hao and Qi (2015), and neighborhood consistency regularization Hao and Qi (2015).

Based on a similar assumption to the above works, Yin et al. (2011) proposed the Associate-Predict model (APM) for frontal face synthesis. The APM model has two steps for estimating the frontal pose patch from a non-frontal pose patch x_{pt}. In the "Associate" step, x_{pt} is associated with the most similar patch in D_p. In the "Predict" step, the associated patch's corresponding patch in D_0 is directly utilized as a prediction for the frontal pose patch of x_{pt}, and the predicted frontal patch is employed for the purpose of face matching.

Unlike many least square regression-based methods for face synthesis, Li et al. (2009) proposed the formulation of CCA as a regressor for frontal face reconstruction. In the training phase, CCA is employed to build the correlation-maximized subspace shared by the frontal and non-frontal patches, as described in Section 4. For frontal face synthesis, the non-frontal patches are first projected into the correlation-maximized subspace. Then, ridge regression is employed to regress them into the frontal face space. Overlapping all the synthesized frontal patches forms the required virtual frontal face. As Li et al. (2009)'s approach is not based on such delicate assumption as that in Chai et al. (2007), they report both higher synthesized image quality and higher recognition performance on the synthesized faces than Chai et al. (2007).

In summary, although the linear regression-based methods are simple, they typically require a certain amount of multi-pose training data. The regressed face images also suffer from the blurring effect and lose critical fine textures for recognition. Moreover, for the approaches based on strict assumptions, e.g., identical local manifold structure across pose, the synthesized faces are not guaranteed to be similar to the person appearing in the input image.

5.3 Nonlinear Regression Models

The appearance variation of the face image across poses is intrinsically nonlinear, as a result of the substantial occlusion and nonlinear warp. To synthesize face images of higher quality, nonlinear regression models have recently been introduced.

The four works reviewed in Section 3 Zhu et al. (2013); Zhang et al. (2013a); Kan et al. (2014); Yim et al. (2015) adopt neural networks as nonlinear regression models for 2D face synthesis, inspired by their power to learn nonlinear transformations. The common point for these four works is that they all first extract pose-robust features, which are then utilized for frontal face recovery. However, they have differences in structure. Zhu et al. (2013); Yim et al. (2015) adopted CNN to extract pose-robust features, and face synthesis is accomplished via a full-connected reconstruction layer. Zhang et al. (2013a) employed single-hidden-layer auto-encoders for frontal face synthesis. Kan et al. (2014) utilized stacked auto-encoders to recover the frontal face image from non-frontal input faces in a progressive manner, thus reducing the difficulty of face synthesis for each auto-encoder. Because of their deep structure, Zhu et al. (2013); Kan et al. (2014); Yim et al. (2015) may synthesize higher quality face images.

The common limitation of Zhu et al. (2013); Zhang et al. (2013a); Kan et al. (2014) is that they are all deterministic networks that recover face images of a fixed pose. In comparison, Zhu et al. (2014a); Yim et al. (2015) designed deep neural networks that can synthesize face images of varied poses. For example, Zhu et al. (2014a) proposed the Multi-View Perceptron (MVP) approach to infer a wide range of pose-varied face images of the same identity, given a single input 2D face. In brief, MVP first extracts pose-robust features with a similar structure as Zhu et al. (2013). The pose-robust feature is then combined with pose selective neurons to generate the reconstruction feature, which is utilized to synthesize faces under the selected pose. The face images of different poses are synthesized with the varied outputs of the pose selective neurons.
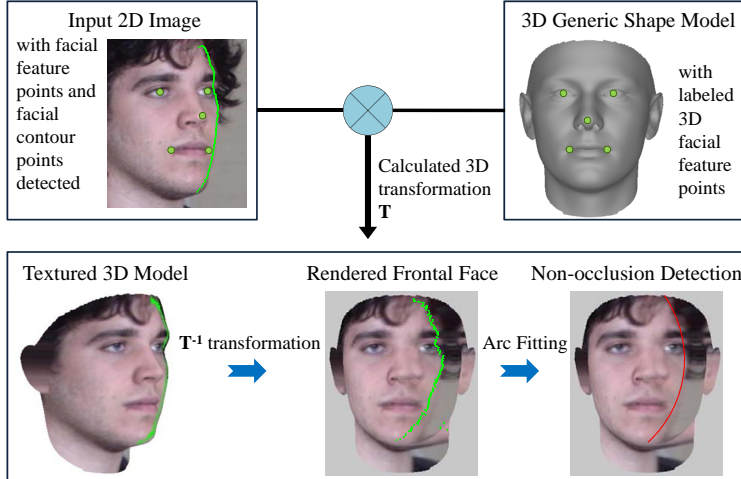
Figure 8: The pipeline for 3D pose normalization from a single face image proposed in Ding et al. (2015). Face regions that are free from occlusion are detected and employed for face recognition.

5.4 Discussion

The 2D pose normalization-based methods conduct face synthesis by calculating the warps across poses for each piece, patch, or pixel in the face image. They require only limited or no training data, and they preserve the fine textures of the original image. With the increase in pose difference, however, the synthesized face image contains more significant stretching artifacts as a result of the dense sampling in relatively small regions. The linear regression model-based methods compensate for this shortcoming but are usually based on strict assumptions about the local manifold structures across pose. They handle a wider range of pose variations and require a moderate amount of training data. However, these assumptions are not guaranteed to hold in practice. The nonlinear regression model-based methods are the most powerful because they model the nonlinear appearance variation caused by pose change. The major shortage is that they require substantial training data and considerable training time. Moreover, the common shortcoming for both the linear or nonlinear regression methods is that the synthesized face image suffers from the blurring effect and loses fine facial textures, e.g., moles, birthmarks, and wrinkles. These peculiarities of textures are crucial for face recognition, as observed by (Park and Jain 2010; Li et al. 2015).

6 Face Synthesis based on 3D Methods

The human head is a complex non-planar 3D structure rotating in 3D space while the face image lies in the 2D domain. The lack of one degree of freedom makes it difficult to conduct face synthesis using 2D techniques alone. The methods reviewed in this section build a 3D model of the human head and then conduct face synthesis based on the 3D face models. Face synthesis using 3D face models is one of the most successful strategies for PIFR. The 3D methods can be classified into three subcategories, i.e., 3D pose normalization from single image, 3D pose normalization from multiple images, and 3D modeling by image reconstruction.

6.1 3D Pose Normalization from Single Image

The 3D pose normalization approaches employ the 3D facial shape model as a tool to correct the nonlinear warp of facial textures appearing in the 2D image. Like the 2D pose normalization methods reviewed in Section 5, they preserve the original pixel values of the input image. As illustrated in Fig. 8, the general principle is that the 2D face image is first aligned with a 3D face model, typically with the help of facial landmarks Jiang et al. (2005); Hassner (2013). Then, the texture of the 2D image is mapped to the 3D model. Lastly, the textured 3D model is rotated to a desired pose and a new 2D image in that pose is rendered. Early approaches utilized simple 3D models, e.g., the cylinder model Gao et al. (2001), the wire

frame model Lee and Ranganath (2003); Zhang et al. (2006), and the ellipsoid model Liu and Chen (2005), to roughly model the 3D structure of the human head, whereas newer approaches strive to build accurate 3D facial shape models.

6.1.1 Normalization using PCA-based Face Models

Most recent approaches build accurate 3D shape models by analyzing a set of aligned 3D facial scans Vetter (1998); Blanz and Vetter (2003). Formally, suppose there are N 3D meshes, each of which is composed of n vertices, and the full vertex-to-vertex correspondence between the N meshes is established such that vertices with the same index in each mesh have the same semantic meaning, e.g., the tip of the nose. The geometry of the ith mesh can be represented as a shape vector that is composed of the 3D coordinates of the n vertices:

$$S_i = (x_1, y_1, z_1, x_2, \cdots, x_n, y_n, z_n)^T \in R^{3n}. \tag{13}$$

The collection of N meshes is represented as $S = [S_1, S_2, \cdots, S_N] \in R^{3n \times N}$. Since semantic correspondence has been established between the 3D scans, it is reasonable to conduct Principle Component Analysis (PCA) on S to model the variations of the 3D facial shape. A new 3D face shape S_t can then be represented as

$$S_t = \bar{S} + A\alpha, \tag{14}$$

where $\bar{S}$ is the mean shape of S. $A \in R^{3n \times h}$ is the matrix stacked of h eigenvectors and $\alpha \in R^h$ contains the coefficients for the eigenvectors.

Given a 2D face image to be normalized, its landmarks are represented as

$$L_t = (\hat{x_1}, \hat{y_1}, \hat{x_2}, \cdots, \hat{x_m}, \hat{y_m})^T \in R^{2m}. \tag{15}$$

Suppose the projection model of the camera is $\mathbf{P}$ which can be the orthogonal projection model or the perspective projection model. $\mathbf{\Omega}(S_t)$ selects m from all its n vertices that correspond to the m facial landmarks in the 2D image. The rigid transformation $\mathbf{T}$ of the face and the shape parameter α are obtained by the following optimization,

$$\min_{T, \alpha} \|L_t - \mathbf{PT\Omega}(S_t)\|_F^2 + \lambda\Phi(\alpha), \tag{16}$$

where $\Phi(\alpha)$ is a regularization term that penalizes the value of α. $\mathbf{T}$ accounts for the translation, rotation, and scaling of the 3D face model S_t. Lastly, the facial texture in the 2D image is mapped to the 3D face model with the shape parameter α and the rigid transformation $\mathbf{T}$. The textured 3D model can be rotated to any desired pose, after which 2D face images under new poses can be rendered by computer graphics tools such as OpenGL.

Although theoretically simple, there are a number of factors to be considered in practical application.

First, **Generic 3D Model** vs **Personalized 3D Model**. Typically, there is only one image for each gallery or probe subject. Inferring a personalized 3D structure from a single 2D image is difficult, therefore approaches that adopt the generic 3D model Ding et al. (2015); Abiantun et al. (2014); Mostafa and Farag (2012) simply calculate $\mathbf{T}$ by setting α as the zero vector, i.e., ignoring the difference between subjects in the 3D structure. This is reasonable to some extent since the difference between subjects in the 3D structure is minor. There are also approaches that emphasize the importance of individual differences in facial structure for face recognition Niinuma et al. (2013); Jo et al. (2015). In this case, the shape parameter α and transformation parameter $\mathbf{T}$ are solved in an alternative optimization fashion. In case of overfitting, Patel and Smith (2009) and Yi et al. (2013) proposed the imposition of regularization on the magnitude of α. Although appealing, these approaches rely on the accurate detection of dense facial landmarks and may suffer from statistical stability problems. This indicates that different values of α are obtained from different images of the same subject, which adversely affects face recognition, as argued by Hassner et al. (2015). Another practical limitation is observed by Jiang et al. (2005), who revealed that the popular 3D face databases contain 3D scans from only a small number of subjects, e.g., 100 in the USF Human ID 3D database Blanz and Vetter (1999), 100 in BU-3DFE database Yin et al. (2008), and 200 in the Basel Face Model Paysan et al. (2009). Therefore, the face space A in Eq. 14 spanned by these scans is quite limited and may not cover the rich variations appearing in testing images.

Second, **Accurate Correspondence**. The accuracy of Eq. 16 depends on the correct correspondence between the 2D landmarks and 3D vertices, i.e., they should be of the same semantic meaning. However, the correspondence between the 2D landmarks on the facial contour and 3D model vertices are pose-dependent, as observed by (Asthana et al. 2011b). This is because the position of facial contour in the 2D image changes along with pose variations. To handle this problem, (Lee et al. 2012; Qu et al. 2014) proposed to detect and discard the moved landmarks. Asthana et al. (2011b) proposed the construction of a lookup table which contains the manually labeled 3D vertices that correspond to the 2D facial landmarks under a set of discrete poses. In the testing phase, the pose of the test image is first estimated and the nearest pose in the lookup table is determined. The corresponding 3D model vertices recorded in the table are employed for Eq. 16. In contrast to Asthana et al. (2011b), (Ding et al. 2012; Zhu et al. 2015) proposed automatic approaches to establish the pose-dependent correspondences online. For example, in Ding et al. (2012), the pose of the test image is first estimated, after which a virtual image under the same pose is rendered using a textured generic 3D model. The facial landmarks of the virtual image are detected. Then, the depth buffer at the position of facial landmarks is searched and corresponding 3D vertices are determined. Compared with Asthana et al. (2011b), Ding et al. (2012)'s approach spares the need for tedious offline labeling work; however, it relies on virtual image rendering and is thus less efficient in the testing phase.

Third, **Occlusion Detection**. The 3D pose normalization approaches rely on texture mapping to synthesize face images in novel poses; however, facial textures are incomplete for a non-frontal face due to self-occlusion. When the 3D model is textured from the non-frontal face, the vertices on the far-side of the 3D model cannot be correctly textured. Therefore, when rendering new face images under the frontal or opposite poses, the facial textures rendered by the occluded 3D vertices are not useful for recognition. To distinguish the occluded and occlusion-free textures in the rendered face image, Li et al. (2014) proposed the use of pose-specific masks in the rendered image. Ding et al. (2012) introduced the Z-buffer algorithm Van Dam and Feiner (2014) to determine the visibility of each vertex on the 3D model. Similarly, Abiantun et al. (2014) proposed the inference of the visibility of each vertex by comparing its normal direction and the viewing direction of the camera. If the angle between the two directions is beyond a certain value, then the vertex is regarded as invisible. Nevertheless, the above methods may not be accurate for occlusion detection because they assume accurate estimation for the pose and 3D facial structure. Recently, inspired by the fact that the major boundary between the occluded and un-occluded regions is the facial contour, Ding et al. (2015) proposed a robust algorithm to detect the facial contour of non-frontal face images. As shown in Fig. 8, the facial contour is projected to the rendered frontal face image and fitted with an arc, which serves as the natural boundary between the occluded and un-occluded facial textures. This strategy detects occlusion precisely and is free from accurate pose estimation.

Fourth, **Synthesis of the Occluded Texture**. After occlusion detection as described above, some approaches employ only the un-occluded textures for recognition Li et al. (2014); Ding et al. (2015), while others synthesize the occluded textures such that a complete face image is obtained. For example, Ding et al. (2012) fill the occluded textures by copying the horizontally symmetrical textures. (Hassner et al. 2015; Zhu et al. 2015) observed that this strategy may produce artifacts when the lighting conditions on both sides of the face are different, and proposed compensation methods, respectively. Also, the human face is asymmetrical, particularly when facial expressions or occlusion are involved, as argued by Abiantun et al. (2014). Instead, Abiantun et al. (2014) proposed to fit the un-occluded pixels to a PCA model that is trained by the frontal faces.

$$\min_c \|x' - \mathbf{\Omega}(Vc + m)\|_F^2 + \lambda\|c\|_1, \tag{17}$$

where x' is the vector containing the un-occluded pixels in the image. V and m are computed by the PCA model with the frontal image training set. $\mathbf{\Omega}$ is a selection matrix that selects the pixels corresponding to x' from $Vc + m$. The learnt PCA coefficients vector c is employed for the reconstruction of the occluded pixels.

Fifth, **Synthesize Frontal Faces** vs **Non-frontal Faces**. The most common setting for the PIFR research and application is that the gallery set is composed of controlled frontal faces, while the probe set contains pose-varied faces. Therefore, there are two distinct choices for face synthesis: (i) Transforming each of the probe images to the frontal pose online Ding et al. (2015); Best-Rowden et al. (2014). In this case, the traditional frontal face recognition algorithms are employed to match each probe image with the gallery images; (ii) Transforming each of the gallery images to a discrete set of pose-varied images offline Niinuma et al. (2013); Hsu and Peng (2013). Given a probe image, its pose is first estimated and the

gallery images of the nearest pose are selected for face matching. This strategy is called the view-based face recognition Beymer (1994). Choice (i) has the advantage of handling continuous pose variations that appear in the probe set, while Choice (ii) divides the pose space into discrete sections and is affected by the accuracy in pose estimation. The latter choice is also less efficient as it stores a large number of gallery images in memory. The advantage of the latter choice, however, lies in the fact that automatic 3D modeling from frontal images is typically easier than modeling from profile faces. This is because dense facial landmark detection, which is important for 3D modeling, is usually unstable for profile faces due to severe self-occlusion.

Sixth, **Handling Expression Variations**. Eq. 14 does not take expression variations into consideration, i.e., the face images are assumed to be of neutral expression. This impacts on recognition in two ways. First, the rendered faces save the expression that appears in the original face image, which may be different from the expression in the image to be compared. To render faces with varied expressions, Jiang et al. (2005) proposed using the MEPG-4-based animation framework to drive the textured 3D face model to present different expressions. Second, the value of the PCA coefficient vector α in Eq. 14 is contaminated by expression variations in the input image. To handle this problem, Chu et al. (2014) proposed expending Eq. 14 as follows.

$$S_t = \bar{S} + [A_{id}\ A_{exp}] \begin{bmatrix} \alpha_{id} \\ \alpha_{exp} \end{bmatrix},$$ (18)

where A_{id} and A_{exp} are PCA models that account for identity and expression variations, respectively. A_{exp} can be trained on databases which contain a set of expressive 3D scans Yin et al. (2008); Cao et al. (2014). Correspondingly, α_{id} and α_{exp} stand for the identity and expression coefficient vectors, respectively. After fitting an image and mapping the facial textures in the input image to the 3D model, the obtained α_{id} remains unchanged, while α_{exp} is forced to be the coefficients of the neutral expression. Therefore, the 3D model and the subsequently rendered images would be of neutral expression. Although it is more powerful, the major drawback of Eq. 18 is that the separation between α_{id} and α_{exp} may be inaccurate.

6.1.2 Normalization using Other Face Models

Apart from modeling 3D face structures with PCA, a number of approaches adopt other 3D modeling methods. Heo (2009) proposed an efficient approach called 3D Generic Elastic Models (GEM) for 3D modeling from a single frontal face image. The fundamental assumption of GEM is that the depth information of human faces does not change dramatically between subjects. It can be approximated from a generic depth map, as long as the frontal face and the depth map are densely aligned. To establish the dense correspondence, 79 facial landmarks are first detected for both the frontal face and the depth map. Delaunay Triangulation is utilized to obtain sparse 2D meshes from the landmarks. To increase the vertex density of the meshes, loop subdivision is employed to generate new vertices using a weighted sum of the existing sparse vertices. As the meshes for the frontal face and depth map are created in the same way, an accurate correspondence is built. The depth value of the frontal mesh vertices is directly borrowed from those of the depth map, while the texture value of the frontal mesh vertices remains the pixel value of the frontal image. Thus, a textured 3D model of the frontal face is obtained. Prabhu et al. (2011) applied GEM to PIFR. They conducted 3D modeling for each frontal gallery image, and then generated non-frontal faces with the same pose as the probe face for matching. GEM has the advantage in efficiency in the 3D modeling phase. It also saves both the original shape and texture information of the frontal face, which is important for face recognition. However, it applies only to the 3D modeling of frontal faces, thus a number of non-frontal faces have to be rendered to extend the gallery set.

GEM is based on the strong approximation of the true depth of the face with a generic depth map. To relieve this assumption, Heo and Savvides (2012b) proposed the Gender and Ethnicity specific GEMs (GE-GEMs) which employs gender and ethnicity specific depth maps for the 3D modeling of frontal faces. The basic assumption of GE-GEMs is that the depth information of faces varies significantly less within the same gender and ethnicity group. The authors empirically show that GE-GEMs can model the 3D shape of frontal faces more accurately than GEM. The shortcoming of GE-GEMs is that it relies on correct gender and ethnicity classification in the testing time, which is accomplished in a semi-automatic manner in this work. Neither GEM or GE-GEM considers the influence of expression variation on the facial depth value. Moeini and Moeini (2015) proposed the Probabilistic Facial Expression Recognition Generic Elastic Model

(PFER-GEM) to generate the image-specific depth map. In PFER-GEM, the image-specific depth map is formulated as the linear combination of depth maps of four typical expressions, therefore the constructed depth map is expression-adaptive. The experimental results show that higher accuracy in 3D face modeling is achieved by PFER-GEM than GEM and GE-GEM.

6.2 3D Pose Normalization from Multiple Images

The approaches reviewed above attempt to conduct 3D modeling from a single input image, since it is the most common setting for real-life face recognition applications. The common shortcoming is that the personalized 3D shape parameters cannot be precisely approximated, since 3D modeling from a single face image is essentially an ill-posed problem. However, in some special applications, e.g., law enforcement, multi-view images are available for each subject during enrollment Zhang et al. (2006, 2008). In this case, it is desirable to utilize the multiple images together to build a more accurate 3D face model Xu et al. (2015).

Zhang et al. (2005, 2008) proposed the Multilevel Quadratic Variation Minimization (MQVM) approach to reconstruct an accurate 3D structure from a pair of frontal and profile face images of a subject. Heo and Savvides (2011) employed the depth information in the profile face image to modify the generic depth map in GEM. In brief, a sparse 3D face shape is constructed by aligning the facial landmarks of the frontal and profile face images. The depth map of the sparse face shape is merged with the generic depth map. The combined depth map replaces the generic depth map in GEM for face synthesis. It is argued that the combined depth map is more accurate and thus more realistic face images can be synthesized. However, there are no experiments to validate the effectiveness of the proposed approach for face recognition.

Similarly, Han and Jain (2012) extended the GEM approach to utilize the complementary information incorporated in the frontal and profile image pair. Their approach is based on Eq. 14 and Eq. 16, by which two 3D models are derived. One is from the frontal face and the other is from the profile face, denoted as S_f and S_p, respectively:

$$S_f = \bar{S} + A\alpha_f,$$
$$S_p = \bar{S} + A\alpha_p. \tag{19}$$

S_f well preserves the 2D shape information of the frontal face, while S_p depicts more accurate depth information. The information in S_f and S_p is complementary since they contain accurate 2D shape information and 3D depth information, respectively. The final 3D shape model for the subject is obtained by replacing the depth value of 3D vertices in S_f by that of S_p. The 3D model is textured by mapping the pixel value in the frontal face image to the 3D model.

The pair of frontal and profile faces is not always available. Mostafa et al. (2012) proposed the reconstruction of a 3D structure of a subject during enrollment from stereo pair images, captured by two cameras from arbitrarily different viewpoints. The stereo pair images are employed to compute a disparity map using a stereo matching algorithm. A cloud of 3D points is estimated from the disparity map and refined by surface fitting, and a 3D triangular mesh is generated from the 3D points as the personalized 3D model for the subject. For recognition, a set of images under discrete poses for each gallery subject is rendered using the 3D model and ray-tracing algorithm. Given a probe image with non-frontal pose, the gallery images of the closest pose are employed for matching. The experimental results highlight that using stereo-based 3D reconstruction to synthesize images is more accurate than using a generic 3D shape model Mostafa et al. (2012); Ali (2014).

6.3 3D Modeling by Image Reconstruction

The 3D pose normalization approaches reviewed in the above two subsections take their cues from a set of facial landmarks for 3D shape reconstruction. Compared to the whole face image, the information contained in the facial landmarks is limited, which creates difficulty for accurate 3D face reconstruction and the subsequent face synthesis. In contrast, the approaches reviewed in this subsection make full use of every pixel in the image to infer the 3D structure of the face image.

Blanz and Vetter (1999, 2003) proposed the 3D Morphable Model (3DMM) approach to simulate the process of image formation by combining a deformable 3D model and computer graphics techniques. The deformable 3D model includes one shape vector and one texture vector. The shape vector S_i for the ith mesh

is identical to Eq. 13 and Eq. 14. Similarly, the texture vector T_i contains the color value of each vertex and is represented as follows,

$$T_i = (r_1, g_1, b_1, r_2, \cdots, r_n, g_n, b_n)^T \in R^{3n}. \tag{20}$$

The collection of the texture vectors for the N scans is represented as $T = [T_1, T_2, \cdots, T_N] \in R^{3n \times N}$. By conducting PCA on T, the texture vector for a new 3D face T_t is represented as

$$T_t = \bar{T} + B\beta, \tag{21}$$

where $\bar{T}$ is the mean texture of T. $B \in R^{3n \times k}$ is the matrix stacked of k eigenvectors and $\beta \in R^k$ contains the coefficients for the eigenvectors.

Computer graphics techniques employed in 3DMM incorporate the Phong illumination model Van Dam and Feiner (2014) and the 3D-to-2D perspective projection model. The illumination model accounts for the wide range of illumination variations in the face image, including cast shadows and specular reflections. Given a single image, 3DMM automatically estimates the 3D shape coefficient vector α, texture coefficient vector β, and parameters of computer graphics models γ by fitting the deformable 3D model with the face image. The model parameters are optimized by a stochastic version of Newton's algorithm, with the objective that the sum of squared differences over all pixels between the rendered image I_{render} and the input image I_{input} should be as similar as possible:

$$\min_{\alpha, \beta, \gamma} \sum_{x, y} \|I_{input}(x, y) - I_{render}(x, y)\|^2. \tag{22}$$

Ideally, the separated shape and texture parameters α and β are only related to identity and thus provide pose and illumination invariance. Face recognition can therefore be conducted by comparing α and β between gallery and probe images. Also, with the optimized α and β, 2D face images can be synthesized under arbitrary poses for the subject appearing in the input image.

The major disadvantage of 3DMM lies in its fitting procedure, which is highly nonlinear and computationally expensive. In practice, the contribution of shape, texture, pose, and illumination parameters to pixel intensity is ambiguous, so the optimization of Eq. 22 is non-convex and prone to be trapped into local minima. Another disadvantage is that the PCA-based texture modelling may hardly represent the fine textures that are particular to individual faces Vetter (1998).

To reduce the computational complexity and relieve the non-convexity of 3DMM, multiple constraints are employed to regularize the fitting procedure in Eq. 22. For example, Romdhani and Vetter (2005) proposed a new fitting algorithm that employs multiple features including pixel intensity, edges, and specular highlights. They utilize a Bayesian framework and maximize the posterior of the shape and texture parameters given the multiple features. The resulting cost function is smoother than Eq. 22 and thus easier to optimize, rendering the fitting procedure more stable. With the proposed fitting algorithm, the authors reported higher face recognition accuracy and significantly better efficiency than the original fitting algorithm in 3DMM. Other constraints incorporate facial symmetry prior Hu et al. (2013), et al.

The non-convex optimization problem in Eq. 22 can also be relieved with various multi-stage strategies. For example, (Kang et al. 2008; Hu et al. 2012) proposed the multi-resolution fitting methods that fit the 3DMM model to the down-sampled low-resolution images and the original high-resolution image, successively. Aldrian and Smith (2013) proposed the sequential estimation of shape and texture parameters. First, the shape parameter α is estimated using Eq. 16, then the texture parameter β and the illumination conditions are estimated by simulating the image formation process. The optimization in each of the two steps is convex and thus can be solved with a global optimum. Besides, Hu et al. (2014) proposed to first remove the illumination component of the input image using illumination normalization operations, and then run 3DMM to optimize the shape, texture, and pose parameters only.

6.4 Discussion

3D pose normalization methods estimate only the pose and shape of a subject from one or multiple 2D face images, while the texture information is directly mapped from the input 2D image to the 3D model. The rendered images from the textured 3D model are consequently realistic in appearance. Their shortcoming is

that they usually make use of limited information, e.g., coordinates of dense facial landmarks, to estimate the pose and shape parameters. The error in pose and shape estimation results in undesirable artifacts in the subsequent texture mapping and face synthesis operations, which adversely affects face recognition. The 3D pose normalization methods also need special operations to handle the issue of missing data caused by self-occlusion. In contrast, image reconstruction-based 3D modeling methods make full use of the textures appearing in the input 2D image. The pose, shape, texture, and illumination parameters are estimated together by reconstructing the 2D image. Face synthesis can be conducted with the estimated shape and texture parameters. Their shortcoming lies in the difficulty in image fitting, which is often a non-convex problem, thus the shape and texture parameters obtained may not be accurate, which results in error in the synthesized face images. As with the 2D face synthesis methods by regression models, the synthesized images also lose detailed textures, e.g., moles, birthmarks, and wrinkles, which is disadvantageous from the perspective of face recognition (Park and Jain 2010; Li et al. 2015). Finally, one common disadvantage of 3D-based face synthesis methods is that they tend to lose information of the background surrounding the face region. Recently, Zhu et al. (2015) proposed to conduct 3D modeling of both the face and its background, so that both information can be saved in the pose-normalized face image.

7 Hybrid Methods

The category of hybrid methods combines one or more of the aforementioned strategies for PIFR, aiming to make use of the complementary advantages of these methods. The hybrid approaches are less studied in the literature but tend to be more powerful than any single PIFR category, and hold more promise for solving real-world PIFR problems. Several successful combinations are reviewed below.

A number of approaches combine pose-robust feature extraction and multi-view subspace learning Prince et al. (2008); Li et al. (2009); Fischer et al. (2012); Annan et al. (2012). Instead of extracting holistic features from the whole face image, they extract pose-robust features around facial landmarks, substantially reducing the difficulty of multi-view subspace learning. It has been shown that this strategy significantly enhances the performance of PIFR systems. Zhu et al. (2014b) proposed a combination of 2D-based face synthesis and pose-robust feature extraction. After obtaining the canonical view of a face image, the respective face component-level features are extracted for face matching. Ding et al. (2015) proposed the combination of 3D-based face synthesis and multi-view subspace learning, inspired by the fact that frontal faces synthesized from non-frontal images of distinct poses differ in image quality, which needs to be improved by multi-view subspace learning.

It is also possible to employ two or more categories of techniques independently and fuse their estimates into a single result. For example, Kim and Kittler (2006) proposed an expert fusion system where pose-robust feature extraction expert, multi-view subspace learning expert, and face synthesis expert run independently. Results of the experts are fused in the score level and impressive performance improvement is observed. Besides, it might be helpful to use 2D pose normalization and 3D pose normalization independently. The 2D pose normalization methods are not good at correcting the nonlinear texture warping caused by pose variation, but they can save all the information in the original 2D image, e.g., facial contour, hairstyle, clothes, and backgrounds Berg and Belhumeur (2012); Taigman et al. (2014), which are also important cues for face recognition Kumar et al. (2009). Conversely, 3D pose normalization methods are good at correcting nonlinear texture warping but tend to lose information outside of the facial area, as 3D face models only incorporate the facial area. Combining the two face synthesis methods might be helpful for achieving a stronger PIFR system. However, to the best of our knowledge, this fusion strategy has not been explored in the literature.

8 Relations of the Four Categories

In the above sections, we have discussed the pose-robust feature extraction methods, multi-view subspace learning methods, and face synthesis-based methods independently. Indeed, the three categories of methods try to solve the PIFR problem from different perspectives, as illustrated in Eq. 1. In this section, we discuss the relative pros and cons of different categories.

High quality face representation is critical for the traditional NFFR problem, and deep learning-based methods achieve great success in representation learning Taigman et al. (2015); Sun et al. (2015); Schroff et al. (2015); Ding and Tao (2015). In the case of PIFR, we expect the pose-robust features extracted by powerful deep models continue to play a critical part, provided the existence of massive labeled multi-pose training data. However, multi-pose training data that is large enough to drive complicated deep models maybe difficult to collect in real-world applications.

It is not easy to independently exploit multi-view subspace learning methods because they are based on the ideal assumption that simple pose-specific projections eliminate the cross-pose difference of faces. Besides, their performance closely depends on the amount of labeled multi-pose training data. In practice, multi-view subspace learning methods should be combined with pose-robust features, which contribute to reducing the cross-pose gap.

Pose normalization-based face synthesis strategies are particularly successful when there is no multi-pose training data or the training data is small. Their main limitation is the artifacts in the synthesized image caused by the inaccurate estimation of facial shape or pose parameters. These artifacts change the original appearance of the subject, and thus deteriorate the subsequently extracted features, causing an adverse impact on high-precision face recognition, as empirically found in a recent work Ding and Tao (2015).

Therefore, both pose-robust feature extraction and pose normalization are promising solutions to PIFR. In practice, the choice of the most appropriate PIFR method mainly depends on the availability and size of multi-pose training data. Besides, the degree of pose variation is another important factor. For near-frontal or half-profile face images, existing pose-robust feature extraction methods have already achieved high accuracy. For profile face images, the synthesis-based approaches maybe useful to bridge the huge appearance gap between poses. Moreover, hybrid methods try to solve the PIFR problem from multiple perspectives; therefore they may be more promising to handle the real-world PIFR problem, as described in Section 7.

9   Performance Evaluations

To compare the performance of different PIFR algorithms on a relatively fair basis, a variety of face datasets have been established that range in scale and popularity. Table 2 summarizes the existing datasets for PIFR research. Of the existing databases, FERET Phillips et al. (2000), CMU-PIE Sim et al. (2003), Multi-PIE Gross et al. (2010), and LFW Huang et al. (2007) are the most widely explored. In the following, we conduct comparison and analysis of existing PIFR approaches based on their performance on the four databases.

9.1   Evaluation on FERET and CMU-PIE

The performances of representative PIFR algorithms on FERET and CMU-PIE are summarized in Table 3 and Table 4, respectively. Since the vast majority of existing works conduct experiments on face identification, we report their rank-1 identification rates as a metric for performance evaluation. Note that the mean accuracy in Table 3 and Table 4 is calculated based on the performance reported in the original papers.

The multi-pose subset of FERET and the CMU-PIE database are described in Table 2. For FERET, the frontal pose images are used as the gallery set, while all the non-frontal images are utilized as probe sets. For CMU-PIE, methods in Table 4 use face images captured under neural expression and normal illumination conditions. The frontal faces are utilized as the gallery set while the rest of the images compose the probe sets. Note that some methods in Table 3 and Table 4 need images of half subjects for training, so the number of gallery subjects is 100 or 34 for these methods.

Asthana et al. (2011b), Li et al. (2012), and Ding et al. (2015) reported the lowest error rates on FERET and CMU-PIE, demonstrating the power of 2D/3D pose normalization approaches. This may be because pose normalization approaches correct the distortion caused by pose variation while saving the valuable texture detail. The performance of these approaches has almost reached saturation point on FERET and CMU-PIE Ding et al. (2015). However, these results maybe unrealistically optimistic as they assume an ideal situation in which the illumination and expression conditions remain the same across poses. It has been shown that the combined pose and illumination variations have a major influence on the performance

of PIFR algorithms on the CMU-PIE database Zhang and Gao (2012); Aldrian and Smith (2013); Kafai et al. (2014).

9.2 Evaluation on Multi-PIE

More recent approaches report recognition performance on the Multi-PIE database, which covers images from more subjects and a wider range of pose, expression, and recording session variations. The recent representative approaches that have conducted experiments on Multi-PIE are tabulated in Table 5.

Several evaluation protocols on Multi-PIE exist, within which the protocol defined by Asthana et al. (2011b) maybe the most popular one in the literature. Under this protocol, images of the first 200 subjects are employed for training, and images of the remaining 137 subjects are used for testing. Faces in the images have neutral expressions and frontal illumination, and were acquired across four recording sessions. The frontal images from the earliest recording sessions for the testing subjects are collected as the gallery set. The non-frontal images of the testing subjects constitute fourteen probe sets. Performance of approaches that adopt this protocol is reported in Table 6. Among the methods in Table 6, Asthana et al. (2011b) reported a mean accuracy of 86.8% for probe sets whose yaw angles are within $\pm45°$ by exploring the 3D pose normalization technique without occlusion detection. Subsequently, Zhu et al. (2013) improved the performance to 95.6% by learning pose-robust features via DNN. Ding et al. (2015) then achieved accuracy of 99.5% by comprehensively employing 3D pose normalization, image-specific occlusion detection, and multi-task discriminative subspace learning.

Considering that the protocol proposed by Asthana et al. (2011b) is relatively simple, Zhu et al. (2014a) proposed extending Asthana et al. (2011b)'s protocol by incorporating probe images under all 20 illumination conditions. The gallery images remain the same as those in Asthana et al. (2011b). Similar to the experiments on CMU-PIE, the performance of the algorithms drops significantly under combined pose and illumination variations. For example, the average performance of Zhu et al. (2013) and Ding et al. (2015) drop to 78.5% and 96.5% for probe sets within $\pm45°$, respectively.

The robustness of PIFR algorithms to the combined pose and expression variations is less evaluated in the literature. (Schroff et al. 2011; Kan et al. 2012; Chu et al. 2014; Zhu et al. 2015) provide the evaluations on the Multi-PIE database, while Jiang et al. (2005) provide an evaluation on combined pose and expression variations on the CMU-PIE database. The experimental results in Jiang et al. (2005); Chu et al. (2014); Zhu et al. (2015) show that by synthesizing neutral and frontal face images with 3D normalization techniques, the performance of traditional face recognition algorithms is significantly improved.

9.3 Evaluation on LFW

The above three most popular datasets for PIFR, i.e., FERET, CMU-PIE, and Multi-PIE, are small in scale and collected under laboratory conditions, which means that they lack the diversity that appears in real-life faces. Some recent PIFR algorithms evaluate their performance on the uncontrolled LFW database, which is composed of face images captured under practical scenarios. However, it is important to note that LFW is actually designed for the NFFR task rather than PIFR, since the yaw values of more than 96% of LFW images are within $\pm30°$.

Performance of PIFR methods on LFW is tabulated in Table 7. The works by Ding et al. (2015); Arashloo and Kittler (2014); Li and Hua (2015) highlight the importance of extracting pose-robust features. In Ding et al. (2015)'s work, both holistic level and facial component level features are extracted from dense facial landmarks. After fusing face representations from both levels, they achieve a $95.58 \pm 0.34\%$ accuracy under the "Unrestricted, Label-Free Outside Data" protocol. In Arashloo and Kittler (2014)'s work, semantically corresponding patches are densely sampled via MRF and features are extracted by multiple face descriptors. By fusing a number of face representations for each image, they achieve a $95.89 \pm 1.94\%$ accuracy under the "Image-Restricted, No Outside Data" protocol.

In comparison, Zhu et al. (2014b); Ding et al. (2015); Hassner et al. (2015); Zhu et al. (2015) show that various face synthesis techniques are helpful to promote the recognition performance on LFW. For example, Zhu et al. (2014b) employed neural networks for 2D based face synthesis, and reported a $96.45 \pm 0.25\%$ verification rate using the "Unrestricted, Labeled Outside Data" protocol. In comparison, Ding et al. (2015); Hassner et al. (2015); Zhu et al. (2015) utilized 3D pose normalization techniques for frontal face synthesis.

Ding et al. (2015) reported a $92.95 \pm 0.37\%$ verification rate under the "Unrestricted, Label-Free Outside Data" protocol, using only facial texture that is visible to both faces to be compared. Hassner et al. (2015) achieved a $91.65 \pm 1.04\%$ verification rate with the "Image-Restricted, Label-Free Outside Data" protocol. With higher-fidelity 3D pose normalization operations, Zhu et al. (2015) reported $95.25 \pm 0.36\%$ accuracy under the "Unrestricted, Label-Free Outside Data" protocol.

It is clear that pose-robust feature extraction and face synthesis are two mainstream strategies to handle real-world pose variations. Besides, as LFW is designed for the NFFR task rather than PIFR, new datasets that include full range of pose-variation images is desired to further test the performance of PIFR approaches.

9.4 Efficiency Comparison of PIFR Algorithms

In this subsection, the efficiency of some recent PIFR algorithms is briefly discussed, so that readers can have a more comprehensive sense of the pros and cons of different approaches. However, it is important to note that the algorithms are diverse in implementation. Therefore, this subsection presents a qualitative comparison rather than a quantitative comparison.

Within pose-robust feature extraction methods, the PAF approach proposed by Yi et al. (2013) costs 70.9 ms (including facial landmark detection) to re-establish the dense semantic correspondence between images on an Intel P4 CPU @ 2.0GHz, and spends 18.7 ms on feature extraction at the semantically corresponding points. In comparison, the MRF matching model by Arashloo and Kittler (2013) is computationally expensive. It takes 1.4 seconds to establish the semantic correspondence between two images of size 112×128 pixels on a NAVIDIA Geforce GTX 460 GPU, while their previous work Arashloo and Kittler (2011) costs more than 5 minutes on a CPU for the same task. Besides, the DNN-based FIP approach Zhu et al. (2013) spends 0.28 seconds to process one 64×64 pixel image on an Intel Core i5 CPU @ 2.6GHz.

The multi-view subspace learning methods are generally fast in the testing time. Their efficiency mainly differs in the training phase. For example, the matlab implementation of the MvDA approach Kan et al. (2012) takes about 120 seconds for model training on an Intel Core i5 CPU @ 2.6GHz, with training data whose dimension is 600 from 4,347 images of 7 poses.

The efficiency of 2D-based face synthesis methods also differ significantly from each other. Piece-wise-warping costs about 0.05 seconds using a single core Intel 2.2GHz CPU Taigman et al. (2014). The MRF-based face synthesis method by Ho and Chellappa (2013) takes less than two minutes for frontal face synthesis from an input face image of size 130×150 pixels, using an Intel Xeon CPU @ 2.13 GHz.

Within the 3D-based face synthesis methods, pose normalization based methods are generally efficient. For example, pose normalization methods based on generic 3D face model cost about 0.1 seconds for frontal face synthesis from a 250×250 pixel color image, using matlab implementation on an Intel Core i5 CPU @ 2.6GHz Ding et al. (2015); Hassner et al. (2015). The matlab implementation of the more complicated HPEN approach Zhu et al. (2015) costs about 0.75 seconds for pose normalization on an Intel Core i5 CPU @ 2.6GHz. In comparison, image reconstruction-based 3D face synthesis methods are computationally expensive. For example, the classical 3DMM method takes 4.5 minutes on a workstation with an Intel P4 CPU @ 2GHz to fit the 3D face model and a 2D image Blanz and Vetter (2003).

10 Summary and Concluding Remarks

Pose-Invariant Face Recognition (PIFR) is the primary stumbling block to realizing the full potential of face recognition as a passive biometric technology. This fundamental human ability poses a great challenge for computer vision systems. The difficulty stems from the immense within-class appearance variations caused by pose change, e.g., self-occlusion, nonlinear texture distortion, and coupled illumination or expression variations.

In this paper, we reviewed representative PIFR algorithms and classified them into four broad categories according to their strategy to bridge the cross-pose gap: pose-robust feature extraction, multi-view subspace learning, face synthesis, and hybrid approaches. The four categories of approach tackle the PIFR problem from distinct perspectives. The pose-robust features can be grouped into two sub-categories: engineered features and learning-based features, depending on whether the feature is extracted by manually designed descriptors or machine learning models. The multi-view subspace learning approaches are divided into

linear methods and nonlinear methods, of which nonlinear methods show more promising performance. The face synthesis category incorporates 2D-based face synthesis methods and 3D-based face synthesis methods, depending on whether synthesis is conducted in the 3D domain or 2D domain. Synthesis can be accomplished by simple pose normalization or various regression models, and the normalization methods have the advantage of retaining the detail of facial textures. Lastly, the hybrid methods combine two or more of these strategies for high performance PIFR.

Viewing the recent progress in PIFR, we have noticed some encouraging progress for each category of methods reviewed. For example, extracting pose-robust features from semantic corresponding patches is becoming easier due to the rapid progress in facial landmark detection. Researchers have also explored more advanced tools, e.g., 3D shape models, stereo matching, MRF, and GMM, to establish much denser semantic correspondence. The recent explosion of nonlinear machine learning models, e.g., deep neural networks, are utilized to extract pose-robust features, learn the multi-view subspace, and synthesize faces under novel poses. Face synthesis based on 3D methods continues to be a hot topic because these methods work directly and explicitly to reduce the cross-pose gap in the 3D domain, without requiring a large amount of multi-pose training data. The 3D-based face synthesis approaches are still far from perfect, and more accurate and stable algorithms are expected to be developed. Another noticeable phenomenon is the increased number of hybrid approaches developed in recent years, which feature high PIFR performance.

Accompanying the development of PIFR algorithms is the introduction to the field of new databases, e.g., Multi-PIE and IJB-A, which allow for more accurate evaluation to be conducted in more challenging environments. However, images in Multi-PIE are collected under laboratory conditions and thus are not representative of realistic scenarios, which may cause the PIFR task to be unrealistically easy. The popular LFW database is designed for NFFR and contains a very limited number of profile or half-profile faces. Fortunately, the newly introduced IJB-A database may fill the gap of unconstrained face database for PIFR research, and developing larger scale unconstrained databases for PIFR is an emergent task. Equally important is the reasonably designed evaluation protocol for each database, so that various algorithms can be directly compared. Once such databases and corresponding evaluation protocols are available and being adopted, the real performance of existing PIFR algorithms in practical scenarios will be revealed, and new insights for PIFR will be inspired.

Although great progress in PIFR has been achieved, there is still much room for improvement, and the performance of existing approaches needs to be further evaluated on real-world databases. To meet the requirement of practical face recognition applications, we propose the following design criteria as a guide for future development.

- **Fully automatic:** The PIFR algorithms should work autonomously, i.e., require no manual facial landmark annotations or pose estimation, etc. The recent progress of profile-to-profile landmark detection may enable this goal realized in the near future Xiong and De la Torre (2015).

- **Full range of pose variation:** The PIFR algorithms should cover the full range of pose variations that might appear in the face image, including the yaw, pitch, and combined yaw and pitch. In particular, recognition of profile faces is very difficult and largely under-investigated. For pose normalization-based methods, the difficulty lies in the larger error of shape and pose estimation for profile faces. For pose-robust feature extraction-based methods, the difficulty is largely due to the lack of training data composed of labeled large-pose faces. Designing more advanced PIFR algorithms and collecting large-pose training data may be equally important.

- **Recognition from a single image:** The PIFR algorithms should be able to recognize a single non-frontal face utilizing a single gallery image per person. This is the most challenging but also the most common setting for real-world applications.

- **Robust to combined facial variations:** As explained in Section 1, the pose variation is often combined with illumination, expression, and image quality variations. A practical PIFR algorithm should also be robust to combined facial appearance variations. Existing pose normalization-based methods are sensitive to facial expression variations, due to the limitation of representation power of existing 3D face models. Therefore, more expressive 3D models that are competent to model non-rigid expression variations are required. Besides, the rich experience to handle uncontrolled illumination and image quality variations in the NFFR task may help tackle the more challenging PIFR task.

- **Matching between two arbitrary poses:** The most common setting for existing PIFR algorithms is to identify non-frontal probe faces from frontal gallery images. However, it is desirable to be able to match two face images with arbitrarily different poses, for both identification and verification tasks. One extreme example is to match a left-profile face to a right-profile face. In this case, the two faces have little visible regions in common. Using facial symmetry is an intuitive solution, i.e., turning the faces to the same direction Maurer and Von Der Malsburg (1996). However, facial symmetry does not exactly hold true for high-resolution images where fine facial textures are clear. For low-resolution face images, e.g., frames of surveillance videos Beveridge et al. (2015), this strategy may work well.

- **Reasonable requirement of labeled training data:** Although a large amount of labeled multi-pose training data helps to promote the performance of pose-robust feature extraction based PIFR algorithms, it is not necessarily available because labeled multi-pose data are difficult to collect in many practical applications. Possible solutions may incorporate making use of 3D shape priors of the human head and combining unsupervised learning algorithms LeCun et al. (2015). Otherwise, methods that do not rely heavily on large amounts of training data, e.g., pose-normalization based methods, are advantageous.

- **Efficient:** The PIFR algorithms should be efficient enough to realize the requirement of practical applications, e.g., video surveillance and digital entertainment. Therefore, approaches that are free from complicated optimization operations in the testing time are preferable.

In the future, we would expect to see the evaluation of fully automatic and efficient PIFR algorithms on newly-developed real-life and large-scale face databases. Since the existing PIFR approaches handle pose variations from distinct perspectives, we would like to see individual improvement in pose-robust feature extraction, multi-view subspace learning, and face synthesis. For example, it is anticipated that the engineered pose-robust features will be extracted from semantic corresponding patches both densely and efficiently. The powerful nonlinear machine learning models, e.g., deep neural networks, are expected to be fully explored, but at the cost of reasonable amount of multi-pose training data. The face synthesis methods are expected to be able to accurately recover facial shape and textures under varied poses, without artifact or statistical stability problems, possibly by investigating more cues in the image or by combining several synthesis strategies. We would also expect that the combination of several advanced techniques from multiple aspects, i.e., novel hybrid solutions, will better accommodate the complex variations that appear in real images.

References

Ramzi Abiantun, Utsav Prabhu, and Marios Savvides. 2014. Sparse Feature Extraction for Pose-Tolerant Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 36, 10 (2014), 2061–2073.

Timo Ahonen, Abdenour Hadid, and Matti Pietikainen. 2006. Face description with local binary patterns: Application to face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 12 (2006), 2037–2041.

Shotaro Akaho. 2006. A kernel method for canonical correlation analysis. *arXiv preprint cs/0609071* (2006).

Oswald Aldrian and William AP Smith. 2013. Inverse rendering of faces with a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 5 (2013), 1080–1093.

Asem Ali. 2014. A 3D-based Pose Invariant Face Recognition at a Distance Framework. *IEEE Trans. Inf. Forensics Security* 9, 12 (2014), 2158–2169.

Galen Andrew, Raman Arora, Jeff Bilmes, and Karen Livescu. 2013. Deep canonical correlation analysis. In *Proc. Int. Conf. Mach. Learn.* 1247–1255.

Li Annan, Shiguang Shan, and Wen Gao. 2012. Coupled bias–variance tradeoff for cross-pose face recognition. *IEEE Trans. Image Process.* 21, 1 (2012), 305–315.

Shervin Rahimzadeh Arashloo and Josef Kittler. 2011. Energy normalization for pose-invariant face recognition based on MRF model image matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 6 (2011), 1274–1280.

Shervin Rahimzadeh Arashloo and Josef Kittler. 2013. Efficient processing of MRFs for unconstrained-pose face recognition. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.* 1–8.

Shervin Rahimzadeh Arashloo and Josef Kittler. 2014. Class-Specific Kernel Fusion of Multiple Descriptors for Face Verification Using Multiscale Binarised Statistical Image Features. *IEEE Trans. Inf. Forensics Security* 9, 12 (2014), 2100–2109.

Shervin Rahimzadeh Arashloo, Josef Kittler, and William J Christmas. 2011. Pose-invariant face recognition by matching on multi-resolution MRFs linked by supercoupling transform. *Comput. Vis. Image Underst.* 115, 7 (2011), 1073–1083.

Ahmed Bilal Ashraf, Simon Lucey, and Tsuhan Chen. 2008. Learning patch correspondences for improved viewpoint invariant face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1–8.

Ahmed Bilal Ashraf, Simon Lucey, and Tsuhan Chen. 2010. Fast image alignment in the Fourier domain. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2480–2487.

Akshay Asthana, Michael J Jones, Tim K Marks, Kinh H Tieu, and Roland Goecke. 2011a. Pose Normalization via Learned 2D Warping for Fully Automatic Face Recognition.. In *Proc. Brit. Mach. Vis. Conf.* 1–11.

Akshay Asthana, Tim K Marks, Michael J Jones, Kinh H Tieu, and M Rohith. 2011b. Fully automatic pose-invariant face recognition via 3D pose normalization. In *Proc. IEEE Int. Conf. Comput. Vis.* 937–944.

Akshay Asthana, Conrad Sanderson, Tamás D Gedeon, and Roland Goecke. 2009. Learning-based Face Synthesis for Pose-Robust Recognition from Single Image.. In *Proc. Brit. Mach. Vis. Conf.* 1–10.

Jeremiah R Barr, Kevin W Bowyer, Patrick J Flynn, and Soma Biswas. 2012. Face recognition from video: A review. *Int. J. Pattern Recognit. Artif. Intell.* 26, 05 (2012).

Thomas Berg and Peter N Belhumeur. 2012. Tom-vs-Pete Classifiers and Identity-Preserving Alignment for Face Verification.. In *Proc. Brit. Mach. Vis. Conf.*, Vol. 2. 7.

Lacey Best-Rowden, Hu Han, Charles Otto, Brendan Klare, and Anil K Jain. 2014. Unconstrained face recognition: Identifying a person of interest from a media collection. *IEEE Trans. Inf. Forensics Security* 9, 12 (2014), 2144–2157.

J Ross Beveridge, Hao Zhang, and others. 2015. Report on the FG 2015 Video Person Recognition Evaluation. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.* 1–8.

David Beymer and Tomaso Poggio. 1995. Face recognition from one example view. In *Proc. IEEE Int. Conf. Comput. Vis.* 500–507.

David J Beymer. 1994. Face recognition under varying pose. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 756–761.

Soma Biswas, Gaurav Aggarwal, Patrick J Flynn, and Kevin W Bowyer. 2013. Pose-robust recognition of low-resolution face images. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 12 (2013), 3037–3049.

Volker Blanz and Thomas Vetter. 1999. A morphable model for the synthesis of 3D faces. In *Proc. Annual Conf. Computer Graphics and Interactive Techniques.* 187–194.

Volker Blanz and Thomas Vetter. 2003. Face recognition based on fitting a 3D morphable model. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 9 (2003), 1063–1074.

Fred L. Bookstein. 1989. Principal warps: Thin-plate splines and the decomposition of deformations. *IEEE Trans. Pattern Anal. Mach. Intell.* 11, 6 (1989), 567–585.

Kevin W Bowyer, Kyong Chang, and Patrick Flynn. 2006. A survey of approaches and challenges in 3D and multi-modal 3D+2D face recognition. *Comput. Vis. Image Underst.* 101, 1 (2006), 1–15.

Roberto Brunelli and Tomaso Poggio. 1993. Face recognition: Features versus templates. *IEEE Trans. Pattern Anal. Mach. Intell.* 15, 10 (1993), 1042–1052.

Xinyuan Cai, Chunheng Wang, Baihua Xiao, Xue Chen, and Ji Zhou. 2013. Regularized latent least square regression for cross pose face recognition. In *Proc. Int. Joint Conf. Artif. Intell.* 1247–1253.

Chen Cao, Yanlin Weng, Shun Zhou, Yiying Tong, and Kun Zhou. 2014. Facewarehouse: a 3d facial expression database for visual computing. *IEEE Trans. Vis. Comput. Graphics* 20, 3 (2014), 413–425.

Zhimin Cao, Qi Yin, Xiaoou Tang, and Jian Sun. 2010. Face recognition with learning-based descriptor. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2707–2714.

Carlos D Castillo and David W Jacobs. 2007. Using stereo matching for 2-d face recognition across pose. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1–8.

Carlos D Castillo and David W Jacobs. 2009. Using stereo matching with general epipolar geometry for 2D face recognition across pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 12 (2009), 2298–2304.

Carlos D Castillo and David W Jacobs. 2011a. Trainable 3D recognition using stereo matching. In *Proc. IEEE Int. Conf. Comput. Vis. Workshop.* 625–631.

Carlos D Castillo and David W Jacobs. 2011b. Wide-baseline stereo for face recognition with large pose variation. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 537–544.

Xiujuan Chai, Shiguang Shan, Xilin Chen, and Wen Gao. 2007. Locally linear regression for pose-invariant face recognition. *IEEE Trans. Image Process.* 16, 7 (2007), 1716–1725.

Dong Chen, Xudong Cao, Fang Wen, and Jian Sun. 2013. Blessing of dimensionality: High-dimensional feature and its efficient compression for face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 3025–3032.

Baptiste Chu, Sami Romdhani, and Liming Chen. 2014. 3D-aided face recognition robust to expression and pose variations. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1907–1914.

Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. 2001. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 6 (2001), 681–685.

Timothy F Cootes, Christopher J Taylor, David H Cooper, and Jim Graham. 1995. Active shape models-Their training and application. *Comput. Vis. Image Underst.* 61, 1 (1995), 38–59.

Changxing Ding, Jonghyun Choi, Dacheng Tao, and Larry S Davis. 2015. Multi-Directional Multi-Level Dual-Cross Patterns for Robust Face Recognition. *IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2462338* (2015).

Changxing Ding and Dacheng Tao. 2015. Robust Face Recognition via Multimodal Deep Face Representation. *IEEE Trans. Multimedia* 17, 11 (2015), 2049–2058.

Changxing Ding, Chang Xu, and Dacheng Tao. 2015. Multi-task Pose-Invariant Face Recognition. *IEEE Trans. Image Process.* 24, 3 (2015), 980–993.

Liu Ding, Xiaoqing Ding, and Chi Fang. 2012. Continuous pose normalization for pose-robust face recognition. *IEEE Sig. Process. Lett.* 19, 11 (2012), 721–724.

Mika Fischer, Hazım Kemal Ekenel, and Rainer Stiefelhagen. 2012. Analysis of partial least squares for pose-invariant face recognition. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.* 331–338.

Hua Gao, Hazım Kemal Ekenel, and Rainer Stiefelhagen. 2009. Pose normalization for local appearance-based face recognition. In *Proc. Int. Conf. Adv. Biometrics.* 32–41.

Wen Gao, Bo Cao, Shiguang Shan, and others. 2008. The CAS-PEAL large-scale Chinese face database and baseline evaluations. *IEEE Trans. Syst., Man, Cybern. A, Syst., Humans* 38, 1 (2008), 149–161.

Yongsheng Gao, MKH Leung, W Wang, and Siu Cheung Hui. 2001. Fast face identification under varying pose from a single 2-D model view. *IEE Proceedings-Vision, Image and Signal Processing* 148, 4 (2001), 248–253.

Athinodoros S. Georghiades, Peter N. Belhumeur, and David Kriegman. 2001. From few to many: Illumination cone models for face recognition under variable lighting and pose. *IEEE Trans. Pattern Anal. Mach. Intell.* 23, 6 (2001), 643–660.

Daniel González-Jiménez and José Luis Alba-Castro. 2007. Toward pose-invariant 2-D face recognition through point distribution models and facial symmetry. *IEEE Trans. Inf. Forensics Security* 2, 3 (2007), 413–429.

Ralph Gross, Iain Matthews, Jeffrey Cohn, Takeo Kanade, and Simon Baker. 2010. Multi-pie. *Image Vis. Comput.* 28, 5 (2010), 807–813.

Yiwen Guo, Xiaoqing Ding, and Jing-Hao Xue. 2014. MiLDA: A Graph Embedding Approach to Multi-view Face Recognition. *Neurocomput.* (2014), 1255–1261.

Hu Han and Anil K Jain. 2012. 3D face texture modeling from uncalibrated frontal and profile images. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.* 223–230.

Yuanhong Hao and Chun Qi. 2015. A Unified Regularization Framework for Virtual Frontal Face Image Synthesis. *IEEE Sig. Process. Lett.* 22, 5 (2015), 559–563.

Tal Hassner. 2013. Viewing real-world faces in 3D. In *Proc. IEEE Int. Conf. Comput. Vis.* 3607–3614.

Tal Hassner, Shai Harel, Eran Paz, and Roee Enbar. 2015. Effective Face Frontalization in Unconstrained Images. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 4295–4304.

Jingu Heo. 2009. *3D Generic Elastic Models for 2D Pose Synthesis and Face Recognition.* Ph.D. Dissertation. Carnegie Mellon Univ.

Jingu Heo and Marios Savvides. 2011. Rapid 3D face modeling using a frontal face and a profile face for accurate 2D pose synthesis. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.* 632–638.

Jingu Heo and Marios Savvides. 2012a. 3-D Generic Elastic Models for Fast and Texture Preserving 2-D Novel Pose Synthesis. *IEEE Trans. Inf. Forensics Security* 7, 2 (2012), 563–576.

Jingu Heo and Marios Savvides. 2012b. Gender and ethnicity specific generic elastic models from a single 2D image for novel 2D pose face synthesis and recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 12 (2012), 2341–2350.

Huy Tho Ho and Rama Chellappa. 2013. Pose-Invariant Face Recognition Using Markov Random Fields. *IEEE Trans. Image Process.* 22, 4 (2013), 1573–1584.

Gee-Sern Hsu and Hsiao-Chia Peng. 2013. Face Recognition across Poses Using a Single 3D Reference Model. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. Workshops.* 869–874.

Guosheng Hu, Chi-Ho Chan, Josef Kittler, and Bill Christmas. 2012. Resolution-Aware 3D Morphable Model. In *Proc. Brit. Mach. Vis. Conf.* 1–10.

Guosheng Hu, Chi Ho Chan, Fei Yan, William Christmas, and Josef Kittler. 2014. Robust face recognition by an albedo based 3D morphable model. In *Proc. IEEE Int. Joint Conf. Biometrics.* 1–8.

Guosheng Hu, Pouria Mortazavian, Josef Kittler, and William Christmas. 2013. A facial symmetry prior for improved illumination fitting of 3D morphable model. In *Proc. IEEE Int. Conf. Biometrics.* 1–6.

Gary B Huang and Erik Learned-Miller. 2014. *Labeled faces in the wild: Updates and new reporting procedures.* Technical Report. University of Massachusetts, Amherst. 14–18 pages.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments.* Technical Report 07-49. University of Massachusetts, Amherst.

Jian Huang, Pong C Yuen, Wen-Sheng Chen, and Jian Huang Lai. 2007. Choosing parameters of kernel subspace LDA for recognition of face images under pose and illumination variations. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on* 37, 4 (2007), 847–862.

Zhiwu Huang, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2013. Cross-view graph embedding. In *Proc. Asian. Conf. Comput. Vis.* 770–781.

Dalong Jiang, Yuxiao Hu, Shuicheng Yan, Lei Zhang, Hongjiang Zhang, and Wen Gao. 2005. Efficient 3D reconstruction for face recognition. *Pattern Recognit.* 38, 6 (2005), 787–798.

Jaeik Jo, Heeseung Choi, Ig-Jae Kim, and Jaihie Kim. 2015. Single-view-based 3D facial reconstruction method robust against pose variations. *Pattern Recognit.* 48, 1 (2015), 73–85.

Mehran Kafai, Le An, and Bir Bhanu. 2014. Reference Face Graph for Face Recognition. *IEEE Trans. Inf. Forensics Security* 9, 12 (2014), 2132–2143.

Meina Kan, Shiguang Shan, Hong Chang, and Xilin Chen. 2014. Stacked Progressive Auto-Encoders (SPAE) for Face Recognition Across Poses. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*

Meina Kan, Shiguang Shan, Haihong Zhang, Shihong Lao, and Xilin Chen. 2012. Multi-view discriminant analysis. In *Proc. Eur. Conf. Comput. Vis.* 808–821.

Bong-Nam Kang, Hyeran Byun, and Daijin Kim. 2008. Multi-resolution 3D morphable models and its matching method. In *Proc. IEEE Int. Conf. Pattern Recognit.* 1–4.

Tae-Kyun Kim and Josef Kittler. 2006. Design and fusion of pose-invariant face-identification experts. *IEEE Trans. Circuits Syst. Video Technol.* 16, 9 (2006), 1096–1106.

Tae-Kyun Kim, Josef Kittler, Hyun-Chul Kim, and Seok-Cheol Kee. 2003. Discriminant analysis by multiple locally linear transformations. In *Proc. Brit. Mach. Vis. Conf.*, Vol. 1.

Brendan F Klare, Emma Taborsky, Austin Blanton, Jordan Cheney, Kristen Allen, Patrick Grother, Alan Mah, Mark Burge, and Anil K Jain. 2015. Pushing the frontiers of unconstrained face detection and recognition: IARPA Janus Benchmark A. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1931–1939.

Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. 2009. Attribute and simile classifiers for face verification. In *Proc. IEEE Int. Conf. Comput. Vis.* 365–372.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. 2015. Deep learning. *Nature* 521, 7553 (2015), 436–444.

Mun Wai Lee and Surendra Ranganath. 2003. Pose-invariant face recognition using a 3D deformable model. *Pattern Recognit.* 36, 8 (2003), 1835–1846.

Youn Joo Lee, Sung Joo Lee, Kang Ryoung Park, Jaeik Jo, and Jaihie Kim. 2012. Single view-based 3D face reconstruction robust to self-occlusion. *EURASIP Journal on Advances in Signal Processing* 2012, 1 (2012), 1–20.

Martin David Levine and Yingfeng Yu. 2006. Face recognition subject to variations in facial expression, illumination and pose using correlation filters. *Comput. Vis. Image Underst.* 104, 1 (2006), 1–15.

Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. 2009. Maximizing intra-individual correlations for face recognition across pose differences. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 605–611.

Annan Li, Shiguang Shan, Xilin Chen, and Wen Gao. 2011. Cross-pose face recognition based on partial least squares. *Pattern Recognit. Lett.* 32, 15 (2011), 1948–1955.

Dong Li, Huiling Zhou, and Kin-Man Lam. 2015. High-Resolution Face Verification Using Pore-Scale Facial Features. *IEEE Trans. Image Process.* 24, 8 (2015), 2317–2327.

Haoxiang Li and Gang Hua. 2015. Hierarchical-PEP Model for Real-world Face Recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 4055–4064.

Haoxiang Li, Gang Hua, Zhe Lin, Jonathan Brandt, and Jianchao Yang. 2013. Probabilistic elastic matching for pose variant face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 3499–3506.

Peng Li, Yun Fu, Umar Mohammed, James H Elder, and Simon JD Prince. 2012. Probabilistic models for inference about identity. *IEEE Trans. Pattern Anal. Mach. Intell.* 34, 1 (2012), 144–157.

Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. 2012. Morphable displacement field based image matching for face recognition across pose. In *Proc. Eur. Conf. Comput. Vis.* 102–115.

Shaoxin Li, Xin Liu, Xiujuan Chai, Haihong Zhang, Shihong Lao, and Shiguang Shan. 2014. Maximal Likelihood Correspondence Estimation for Face Recognition Across Pose. *IEEE Trans. Image Process.* 23, 10 (2014), 4587–4600.

Qianli Liao, Joel Z Leibo, and Tomaso Poggio. 2013b. Learning invariant representations and applications to face verification. In *Proc. Adv. Neural Inf. Process. Syst.* 3057–3065.

Shengcai Liao, Anil K Jain, and Stan Z Li. 2013a. Partial face recognition: Alignment-free approach. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 5 (2013), 1193–1205.

Danny Little, Sreekar Krishna, A John Jr, and Sethuraman Panchanathan. 2005. A Methodology for Evaluating Robustness of Face Recognition Algorithms with Respect to Variations in Pose Angle and Illumination Angle.. In *Proc. IEEE Int. Conf. Acoustics, Speech, and Signal Processing*, Vol. 2. 89–92.

Chengjun Liu. 2004. Gabor-based kernel PCA with fractional power polynomial models for face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 5 (2004), 572–581.

Tongliang Liu and Dacheng Tao. 2015. Classification with Noisy Labels by Importance Reweighting. *IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2456899* (2015).

Xiaoming Liu and Tsuhan Chen. 2005. Pose-robust face recognition using geometry assisted probabilistic modeling. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 1. 502–509.

David G Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (2004), 91–110.

Thomas Maurer and Christoph Von Der Malsburg. 1996. Single-view based recognition of faces rotated in depth. In *Proc. IEEE Int. Workshop Automatic Face and Gesture Recognit.* 176–181.

Ali Moeini and Hossein Moeini. 2015. Real-World and Rapid Face Recognition towards Pose and Expression Variations via Feature Library Matrix. *IEEE Trans. Inf. Forensics Security* (2015).

Eslam Mostafa, Asem Ali, Naif Alajlan, and Aly Farag. 2012. Pose invariant approach for face recognition at distance. In *Proc. Eur. Conf. Comput. Vis.* 15–28.

Eslam A Mostafa and Aly A Farag. 2012. Dynamic weighting of facial features for automatic pose-invariant face recognition. In *Proc. IEEE Conf. Comput. Robot. Vis.* 411–416.

Marco K Müller, Alexander Heinrichs, Andreas HJ Tewes, Achim Schäfer, and Rolf P Würtz. 2007. Similarity rank correlation for face recognition under unenrolled pose. In *Advances in biometrics*. Springer, 67–76.

Koichiro Niinuma, Hu Han, and Anil K Jain. 2013. Automatic multi-view face recognition via 3D model based pose regularization. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.* 1–8.

Shuxin Ouyang, Timothy Hospedales, Yi-Zhe Song, and Xueming Li. 2014. A Survey on Heterogeneous Face Recognition: Sketch, Infra-red, 3D and Low-resolution. *arXiv preprint arXiv:1409.5114* (2014).

Unsang Park and Anil K Jain. 2010. Face matching and retrieval using soft biometrics. *IEEE Trans. Inf. Forensics Security* 5, 3 (2010), 406–415.

Ankur Patel and William AP Smith. 2009. 3D morphable face models revisited. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1327–1334.

Pascal Paysan, Reinhard Knothe, Brian Amberg, Sami Romdhani, and Thomas Vetter. 2009. A 3D face model for pose and illumination invariant face recognition. In *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance.* 296–301.

Alex Pentland, Baback Moghaddam, and Thad Starner. 1994. View-based and modular eigenspaces for face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 84–91.

P Jonathon Phillips, Hyeonjoon Moon, Syed A Rizvi, and Patrick J Rauss. 2000. The FERET evaluation methodology for face-recognition algorithms. *IEEE Trans. Pattern Anal. Mach. Intell.* 22, 10 (2000), 1090–1104.

Utsav Prabhu, Jingu Heo, and Marios Savvides. 2011. Unconstrained pose-invariant face recognition using 3D generic elastic models. *IEEE Trans. Pattern Anal. Mach. Intell.* 33, 10 (2011), 1952–1961.

Simon JD Prince and James H Elder. 2005. Creating invariance to" nuisance parameters" in face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 2. 446–453.

Simon JD Prince, J Warrell, James H Elder, and Fatima M Felisberti. 2008. Tied factor analysis for face recognition across large pose differences. *IEEE Trans. Pattern Anal. Mach. Intell.* 30, 6 (2008), 970–984.

Chengchao Qu, Eduardo Monari, Tobias Schuchert, and Jurgen Beyerer. 2014. Fast, robust and automatic 3D face model reconstruction from videos. In *Proc. IEEE Int. Conf. Advanced Video and Signal Based Surveillance.* 113–118.

Shervin Rahimzadeh Arashloo and Josef Kittler. 2014. Fast pose invariant face recognition using super coupled multiresolution Markov Random Fields on a GPU. *Pattern Recognit. Lett.* 48 (2014), 49–59.

Sami Romdhani and Thomas Vetter. 2005. Estimating 3D shape and texture using pixel intensity, edges, specular highlights, texture constraints and a prior. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Vol. 2. 986–993.

Jan Rupnik and John Shawe-Taylor. 2010. Multi-view canonical correlation analysis. In *Conf. Data Mining Data Warehouses.* 1–4.

Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: A unified embedding for face recognition and clustering. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 815–823.

Florian Schroff, Tali Treibitz, David Kriegman, and Serge Belongie. 2011. Pose, illumination and expression invariant pairwise face-similarity measure via doppelgänger list comparison. In *Proc. IEEE Int. Conf. Comput. Vis.* 2494–2501.

Abhishek Sharma, Murad Al Haj, Jonghyun Choi, Larry S Davis, and David W Jacobs. 2012a. Robust pose invariant face recognition using coupled latent space discriminant analysis. *Comput. Vis. Image Underst.* 116, 11 (2012), 1095–1110.

Abhishek Sharma and David W Jacobs. 2011. Bypassing synthesis: PLS for face recognition with pose, low-resolution and sketch. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 593–600.

Abhishek Sharma, Abhishek Kumar, H Daume, and David W Jacobs. 2012b. Generalized multiview analysis: A discriminative latent space. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2160–2167.

Terence Sim, Simon Baker, and Maan Bsat. 2003. The CMU pose, illumination, and expression database. *IEEE Trans. Pattern Anal. Mach. Intell.* 25, 12 (2003), 1615–1618.

Karen Simonyan, Omkar M Parkhi, Andrea Vedaldi, and Andrew Zisserman. 2013. Fisher vector faces in the wild. In *Proc. Brit. Mach. Vis. Conf.*, Vol. 1. 1–12.

Richa Singh, Mayank Vatsa, Arun Ross, and Afzel Noore. 2007. A mosaicing scheme for pose-invariant face recognition. *IEEE Trans. Syst., Man, Cybern. B, Cybern.* 37, 5 (2007), 1212–1225.

Yi Sun, Xiaogang Wang, and Xiaoou Tang. 2015. Deeply learned face representations are sparse, selective, and robust. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2892–2900.

Yaniv Taigman, Lior Wolf, Tal Hassner, and others. 2009. Multiple One-Shots for Utilizing Class Label Information.. In *Proc. Brit. Mach. Vis. Conf.* 1–12.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1701–1708.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lior Wolf. 2015. Web-scale training for face identification. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2746–2754.

Xiaoyang Tan, Songcan Chen, Zhi-Hua Zhou, and Fuyan Zhang. 2006. Face recognition from a single image per person: A survey. *Pattern Recognit.* 39, 9 (2006), 1725–1745.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. 2007. General tensor discriminant analysis and gabor features for gait recognition. *IEEE Trans. Pattern Anal. Mach. Intell.* 29, 10 (2007), 1700–1715.

Dacheng Tao, Xuelong Li, Xindong Wu, and Stephen J Maybank. 2009. Geometric mean for subspace selection. *IEEE Trans. Pattern Anal. Mach. Intell.* 31, 2 (2009), 260–274.

Dacheng Tao, Xiaoou Tang, Xuelong Li, and Xindong Wu. 2006. Asymmetric bagging and random subspace for support vector machines-based relevance feedback in image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.* 28, 7 (2006), 1088–1099.

Joshua B Tenenbaum, Vin De Silva, and John C Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 5500 (2000), 2319–2323.

Matthew Turk and Alex Pentland. 1991. Eigenfaces for recognition. *J. Cognitive Neurosci.* 3, 1 (1991), 71–86.

Andries Van Dam and Steven K Feiner. 2014. *Computer graphics: Principles and practice.* Pearson Education.

Thomas Vetter. 1998. Synthesis of novel views from a single face image. *Int. J. Comput. Vis.* 28, 2 (1998), 103–116.

Paul Viola and Michael J Jones. 2004. Robust real-time face detection. *Int. J. Comput. Vis.* 57, 2 (2004), 137–154.

Nannan Wang, Xinbo Gao, Dacheng Tao, and Xuelong Li. 2014b. Facial Feature Point Detection: A Comprehensive Survey. *arXiv preprint arXiv:1410.1037* (2014).

Wen Wang, Zhen Cui, Hong Chang, Shiguang Shan, and Xilin Chen. 2014a. Deeply Coupled Auto-encoder Networks for Cross-view Classification. *arXiv preprint arXiv:1402.2031* (2014).

Renliang Weng, Jiwen Lu, Junlin Hu, Gao Yang, and Yap-Peng Tan. 2013. Robust Feature Set Matching for Partial Face Recognition. In *Proc. IEEE Int. Conf. Comput. Vis.* 601–608.

Laurenz Wiskott, J-M Fellous, N Kuiger, and Christoph Von Der Malsburg. 1997. Face recognition by elastic bunch graph matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7 (1997), 775–779.

John Wright and Gang Hua. 2009. Implicit elastic matching with random projections for pose-variant face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1502–1509.

Xuehan Xiong and Fernando De la Torre. 2015. Global Supervised Descent Method. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 2664–2673.

Chang Xu, Dacheng Tao, and Chao Xu. 2015. Multi-view intact space learning. *IEEE Trans. Pattern Anal. Mach. Intell., doi: 10.1109/TPAMI.2015.2417578* (2015).

Dong Yi, Zhen Lei, and Stan Z Li. 2013. Towards pose robust face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 3539–3545.

Junho Yim, Heechul Jung, ByungIn Yoo, Changkyu Choi, Dusik Park, and Junmo Kim. 2015. Rotating Your Face Using Multi-task Deep Neural Network. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 676–684.

Lijun Yin, Xiaochen Chen, Yi Sun, Tony Worm, and Michael Reale. 2008. A high-resolution 3D dynamic facial expression database. In *Proc. IEEE Int. Conf. Automatic Face and Gesture Recognit.* 1–6.

Qi Yin, Xiaoou Tang, and Jian Sun. 2011. An associate-predict model for face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 497–504.

Haichao Zhang, Yanning Zhang, and Thomas S Huang. 2013b. Pose-robust face recognition via sparse representation. *Pattern Recognit.* 46, 5 (2013), 1511–1521.

Xiaozheng Zhang and Yongsheng Gao. 2009. Face recognition across pose: A review. *Pattern Recognit.* 42, 11 (2009), 2876–2896.

Xiaozheng Zhang and Yongsheng Gao. 2012. Heterogeneous specular and diffuse 3-D surface approximation for face recognition across pose. *IEEE Trans. Inf. Forensics Security* 7, 2 (2012), 506–517.

Xiaozheng Zhang, Yongsheng Gao, and Maylor KH Leung. 2005. Multilevel quadratic variation minimization for 3D face modeling and virtual view synthesis. In *Proc. IEEE Int. Conf. Multimedia Modelling.* 132–138.

Xiaozheng Zhang, Yongsheng Gao, and Maylor KH Leung. 2006. Automatic texture synthesis for face recognition from single views. In *Proc. IEEE Int. Conf. Pattern Recognit.*, Vol. 3. 1151–1154.

Xiaozheng Zhang, Yongsheng Gao, and Maylor KH Leung. 2008. Recognizing rotated faces from frontal and side views: An approach toward effective use of mugshot databases. *IEEE Trans. Inf. Forensics Security* 3, 4 (2008), 684–697.

Xiaozheng Zhang, Yongsheng Gao, and Bai-ling Zhang. 2006. Recognizing Rotated Faces from Two Orthogonal Views in Mugshot Databases. In *Proc. IEEE Int. Conf. Pattern Recognit.*, Vol. 1. 195–198.

Yizhe Zhang, Ming Shao, Edward K Wong, and Yun Fu. 2013a. Random Faces Guided Sparse Many-to-One Encoder for Pose-Invariant Face Recognition. In *Proc. IEEE Int. Conf. Comput. Vis.* 2416–2423.

Sanqiang Zhao and Yongsheng Gao. 2009. Textural Hausdorff distance for wider-range tolerance to pose variation and misalignment in 2D face recognition. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 1629–1634.

Wenyi Zhao, Rama Chellappa, P Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *Comput. Surveys* 35, 4 (2003), 399–458.

Xiangyu Zhu, Zhen Lei, Junjie Yan, Dong Yi, and Stan Z Li. 2015. High-Fidelity Pose and Expression Normalization for Face Recognition in the Wild. In *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.* 787–796.

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2013. Deep learning identity-preserving face space. In *Proc. IEEE Int. Conf. Comput. Vis.* 113–120.

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2014a. Multi-view perceptron: A deep model for learning face identity and view representations. In *Proc. Adv. Neural Inf. Process. Syst.* 217–225.

Zhenyao Zhu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2014b. Recover canonical-view faces in the wild with deep neural networks. *arXiv preprint arXiv:1404.3543* (2014).

Xuan Zou, Josef Kittler, and Kieron Messer. 2007. Illumination invariant face recognition: A survey. In *Proc. IEEE Int. Conf. Biometrics, Theory, Appl. Syst.* 1–8.

Table 1: Taxonomy of Pose-Invariant Face Recognition Approaches

Category	Representative Works
Pose-Robust Feature Extraction	
• Engineered Features	Elastic Bunch Graph Matching Wiskott et al. (1997), Stereo Matching Castillo and Jacobs (2009)
• Learning-based Features	Deep Neutral Networks Zhu et al. (2013); Kan et al. (2014)
Multi-view Subspace Learning	
• Linear Models	CCA Li et al. (2009), PLS Sharma and Jacobs (2011), GMLDA Sharma et al. (2012b), Tied Factor Analysis Prince et al. (2008)
• Nonlinear Models	Kernel CCA Akaho (2006), Deep CCA Andrew et al. (2013)
Face Synthesis based on 2D Methods	
• 2D Pose Normalization	Piece-wise Warping Cootes et al. (2001), Patch-wise Affine Warping Ashraf et al. (2008), MRFs Ho and Chellappa (2013)
• Linear Regression Models	LLR Chai et al. (2007), CCA Li et al. (2009)
• Nonlinear Regression Models	Deep Neutral Networks Zhu et al. (2013); Kan et al. (2014)
Face Synthesis based on 3D Methods	
• Pose Normalization from Single Image	3D Face Shape Model Jiang et al. (2005); Asthana et al. (2011a), Generic Elastic Models Heo (2009); Heo and Savvides (2012b)
• Pose Normalization from Multiple Images	Frontal & Profile Face Pairs Han and Jain (2012), Stereo Matching Mostafa et al. (2012)
• 3D Modeling by Image Reconstruction	3DMM Blanz and Vetter (1999); Aldrian and Smith (2013)
Hybrid Methods	
• Feature & Multi-view Subspace	Block Gabor+PLS Fischer et al. (2012)
• Synthesis & Multi-view Subspace	PBPR-MtFTL Ding et al. (2015)
• Synthesis & Feature	Expert Fusion Kim and Kittler (2006), FR+FCN Zhu et al. (2014b)

Table 2: Standard Datasets for Pose-Invariant Face Recognition

FERET The multi-pose subset of the FERET database Phillips et al. (2000) contains 1,800 face images for 200 subjects across 9 poses. The 9 poses span $-65°$ to $+65°$ in yaw with $16°$ intervals and there is no pitch variation. There is only one image for each subject under a certain pose. The database measures the influence of pose variation to face recognition exclusively, i.e., other influential factors, e.g., illumination, expression, and background, remain the same across poses.

CMU-PIE The CMU Pose, Illumination, and Expression (PIE) database Sim et al. (2003) contains 41,368 face images of 68 subjects across 13 poses, 43 illumination conditions, and 4 different expressions. The 13 pose types contain 9 poses with only yaw variation from the left profile to the right profile, with neighboring poses about $16°$ apart. There are two poses containing pure pitch variations, and another two poses containing combined yaw and pitch variations. All images were captured in a single recording session.

Multi-PIE The Multi-PIE database Gross et al. (2010) contains images of 337 subjects from 15 different viewpoints, 19 illumination conditions, and up to 6 expression types. Multi-PIE covers significantly more subjects than CMU-PIE, and the images of each subject were collected in up to four different recording sessions. There are 13 poses containing only yaw variations, ranging from $-90°$ to $+90°$, spaced in $15°$ intervals. The remaining two poses contain hybrid yaw and pitch variations.

CAS-PEAL-R1 The multi-pose subset of the CAS-PEAL-R1 database Gao et al. (2008) contains 21,832 face images of 1,040 subjects across 21 poses. All images were captured under ambient illumination and neutral expression. Each subject has only one image in each pose. The 21 poses are sampled from the pose space with 7 discrete yaw values and 3 discrete pitch values. The yaw values range from $-67°$ to $+67°$, while the pitch values are approximately $-30°$, $0°$, and $+30°$. Images of the 21 poses are divided into three probe sets, namely PU, PM, and PD, each of which contains images of one particular pitch value.

FacePix The multi-pose subset of the FacePix database Little et al. (2005) contains 181 images for each of the 30 subjects, covering yaw angle variations from $-90°$ to $+90°$, with an interval of $1°$ between nearby poses. There are no pose variations in pitch and the facial expression stays neutral. The illumination condition is simulated ambient lighting and remains the same across poses.

LFW The LFW database contains 13,233 images of 5,749 subjects, of which 1,680 subjects have two or more images. The images in LFW exhibit diverse variations in expression, illumination, and image quality that appear in daily life. However, faces in LFW are detected automatically by the simple Viola-Jones face detector Viola and Jones (2004), which constrains the pose range of faces in LFW. In fact, the yaw values of more than 96% of LFW images are within $\pm 30°$, making LFW a less explored database for PIFR research.

IJB-A The IARPA Janus Benchmark A (IJB-A) database Klare et al. (2015) is a newly published face database which contains 5,712 face images and 2,085 videos from 500 subjects. Similar to LFW, images in IJB-A database are collected from Internet. The key characteristic of IJB-A is that both face detection and facial feature point detection are accomplished manually. Therefore, face images in IJB-A database cover full range of pose variations.

Table 3: Evaluation Summary of Different Categories of PIFR Algorithms on FERET

Publication	Pose Estimation	Landmark Detection	Subject Number	Pose Range (Yaw)	Mean Accuracy
Zhao and Gao (2009)	N/A	N/A	200	$\pm 45°$	73.6
Arashloo and Kittler (2013)	N/A	N/A	200	$\pm 65°$	97.4
Yi et al. (2013)	Auto	Auto	200	$\pm 65°$	97.4
Kan et al. (2014)	N/A	Manual	100	$\pm 65°$	92.5
Li et al. (2009)	Manual	Manual	100	$\pm 65°$	81.5
Sharma et al. (2012a)	Manual	Manual	100	$\pm 65°$	86.4
Ashraf et al. (2008)	Manual	Manual	100	$\pm 65°$	72.9
Annan et al. (2012)	Manual	Manual	100	$\pm 65°$	92.6
Li et al. (2012)	Manual	Manual	100	$\pm 65°$	99.5
Li et al. (2014)	Manual	Manual	100	$\pm 65°$	99.1
Ho and Chellappa (2013)	N/A	Auto	200	$\pm 45°$	95.5
Blanz and Vetter (2003)	Auto	Manual	194	$\pm 65°$	95.8
Asthana et al. (2011b)	Auto	Auto	200	$\pm 45°$	95.6
Mostafa and Farag (2012)	Auto	Auto	200	$\pm 45°$	94.2
Moeini and Moeini (2015)	Auto	Auto	200	$\pm 65°$	99.1
Ding et al. (2015)	Manual	Manual	100	$\pm 65°$	99.6

Mean accuracy is calculated based on the performance reported in the original papers. Methods are organized in the order of pose-robust feature extraction, multi-view subspace learning, 2D-based face synthesis, 3D-based face synthesis, and hybrid categories.

Table 4: Evaluation Summary of Different Categories of PIFR Algorithms on CMU-PIE

Publication	Pose Estimation	Landmark Detection	Subject Number	Pose Range (Yaw/Pitch)	Mean Accuracy
Arashloo and Kittler (2011)	N/A	N/A	68	$\pm 65°$ / $\pm 11°$	92.3
Yi et al. (2013)	Auto	Auto	68	$\pm 65°$ / $\pm 11°$	95.3
Prince et al. (2008)	Manual	Manual	34	$\pm 65°$ / -	-
Sharma and Jacobs (2011)	Manual	Manual	34	$\pm 65°$ / $\pm 11°$	93.9
Chai et al. (2007)	Manual	Manual	68	$\pm 32°$ / $\pm 11°$	94.6*
Asthana et al. (2011a)	Auto	Auto	68	$\pm 32°$ / $\pm 11°$	95.0
Annan et al. (2012)	Manual	Manual	34	$\pm 65°$ / $\pm 11°$	90.9
Li et al. (2012)	Manual	Manual	68	$\pm 65°$ / -	98.4*
Li et al. (2014)	Manual	Manual	68	$\pm 65°$ / -	97.0*
Ho and Chellappa (2013)	N/A	Auto	68	$\pm 45°$ / $\pm 11°$	98.8
Asthana et al. (2011b)	Auto	Auto	67	$\pm 32°$ / $\pm 11°$	99.0
Mostafa and Farag (2012)	Auto	Auto	68	$\pm 45°$ / $\pm 11°$	99.3
Moeini and Moeini (2015)	Auto	Auto	68	$\pm 65°$ / $\pm 11°$	98.2
Ding et al. (2015)	Manual	Manual	68	$\pm 65°$ / $\pm 11°$	99.9*

"*" indicates side information is utilized for model training. Mean accuracy is calculated based on the performance reported in the original papers.

Table 5: Summary of Representative PIFR Algorithms Evaluated on Multi-PIE

Publication	Approach	Pose Estimation	Landmark Detection	Pose Range (Yaw/Pitch)
Wright and Hua (2009)	Implicit Matching	N/A	Auto	±30° / -
Schroff et al. (2011)	Doppelgänger list	Manual	Auto	±90° / ±30°
Zhang et al. (2013a)	RF-SME	N/A	Manual	±75° / -
Zhu et al. (2013)	FIP+LDA	N/A	Manual	±45° / -
Kafai et al. (2014)	RFG	N/A	N/A	±90° / -
Kan et al. (2014)	SPAE	N/A	Auto	±45° / -
Zhu et al. (2014a)	MVP+LDA	N/A	Manual	±60° / -
Sharma et al. (2012a)	ADMCLS	Manual	Manual	±90° / ±30°
Sharma et al. (2012b)	GMLDA	Manual	Manual	±75° / -
Kan et al. (2012)	MvDA	Manual	Manual	±45° / -
Annan et al. (2012)	Ridge Regression	Manual	Manual	±90° / ±30°
Li et al. (2012)	SA-EGFC	Auto	Auto	±45° / -
Li et al. (2014)	MDF-PM	Manual	Manual	±90° / -
Ho and Chellappa (2013)	MRFs	N/A	Auto	±45° / -
Zhu et al. (2013)	RL+LDA	N/A	Manual	±45° / -
Asthana et al. (2011b)	VAAM	Auto	Auto	±45° / -
Prabhu et al. (2011)	GEM	Auto	Auto	±60° / -
Heo and Savvides (2012b)	GE-GEM	Manual	Auto	±30° / -
Zhu et al. (2015)	HPEN+LDA	Auto	Auto	±45° / -
Fischer et al. (2012)	Block Gabor+PLS	Manual	Manual	±90° / ±30°
Ding et al. (2015)	PBPR-MtFTL	Manual	Manual	±90° / ±30°

Table 6: Identification Rates of PIFR Methods on Multi-PIE Using Protocol Defined in Asthana et al. (2011b)

Publication	080 −45°	130 −30°	140 −15°	050 +15°	041 +30°	190 +45°	Mean ≤ ±45°	Mean ≤ ±90°
Zhu et al. (2013)	95.60	98.50	100.0	99.30	98.50	97.80	98.28	-
Kan et al. (2014)	84.90	92.60	96.30	95.70	94.30	84.40	91.37	-
Kafai et al. (2014)	86.40	91.20	96.00	96.10	90.90	85.40	91.00	-
Zhu et al. (2014a)	93.40	100.0	100.0	100.0	99.30	95.60	98.05	-
Kan et al. (2012)	69.67	83.33	93.33	93.00	85.33	69.33	82.33	-
Li et al. (2012)	93.00	98.70	99.70	99.70	98.30	93.60	97.17	-
Ho and Chellappa (2013)	86.30	89.70	91.70	91.00	89.00	85.70	88.90	-
Li et al. (2014)	90.00	94.30	95.30	94.70	93.70	87.70	92.62	-
Asthana et al. (2011b)	74.10	91.00	95.70	95.70	89.50	74.80	86.80	-
Zhu et al. (2015)	97.40	99.50	99.50	99.70	99.00	96.70	98.60	-
Ding et al. (2015)	98.67	100.0	100.0	100.0	100.0	98.33	99.50	92.04

Performance of Kan et al. (2012) is obtained using the code released by the authors, with DCP feature extracted from uniformly divided image regions as the face representation. Performance can be improved when combined with pose-robust feature.

Table 7: Summary of Representative PIFR Algorithms Evaluated on LFW

Publication	Approach	Category	Protocol	Accuracy± Std(%)
Arashloo and Kittler (2013)	MRF-MLBP	Feature	Protocol1	79.08±0.14
Li and Hua (2015)	POP-PEP	Feature	Protocol1	91.10±1.47
Arashloo and Kittler (2014)	MRF-Fusion-CSKDA	Feature	Protocol1	95.89±1.94
Yi et al. (2013)	PAF	Feature	Protocol2	87.77±0.51
Hassner et al. (2015)	Sub-SML+LFW3D	3D Synthesis	Protocol2	91.65±1.04
Zhu et al. (2015)	HPEN+HD-Gabor+JB	3D Synthesis	Protocol2	92.80±0.47
Chen et al. (2013)	High-dim LBP	Feature	Protocol3	93.18±1.07
Ding et al. (2015)	MDML-DCPs	Feature	Protocol3	95.58±0.34
Ding et al. (2015)	PBPR+PLDA	3D Synthesis	Protocol3	92.95±0.37
Zhu et al. (2015)	HPEN+HD-Gabor+JB	3D Synthesis	Protocol3	95.25±0.36
Cao et al. (2010)	Multiple LE+comp	Feature	Protocol4	84.45±0.46
Chen et al. (2013)	High-dim LBP	Feature	Protocol4	95.17±1.13
Yin et al. (2011)	Associate-Predict	2D Synthesis	Protocol4	90.57±0.56
Berg and Belhumeur (2012)	Tom-vs-Pete+Attribute	2D Synthesis	Protocol4	93.30±1.28
Zhu et al. (2014b)	FR+FCN	Hybrid	Protocol4	96.45±0.25

"Feature" stands for the pose-robust feature extraction category. "2D Synthesis" stands for the 2D-based face synthesis category. "3D Synthesis" stands for the 3D-based face synthesis category. "Protocol1" stands for the "Image-Restricted, No Outside Data" protocol Huang and Learned-Miller (2014). "Protocol2" stands for the "Image-Restricted, Label-Free Outside Data" protocol. "Protocol3" stands for the "Unrestricted, Label-Free Outside Data" protocol. "Protocol4" stands for the "Unrestricted, Labeled Outside Data" protocol. Results are directly cited from the original papers.