Maximal Likelihood Correspondence Estimation for Face Recognition Across Pose

Shaoxin Li, *Student Member, IEEE*, Xin Liu, Xiujuan Chai, *Member, IEEE*, Haihong Zhang, *Member, IEEE*, Shihong Lao, *Member, IEEE*, and Shiguang Shan, *Member, IEEE*

***Abstract*— Due to the misalignment of image features, the performance of many conventional face recognition methods degrades considerably in across pose scenario. To address this problem, many image matching-based methods are proposed to estimate semantic correspondence between faces in different poses. In this paper, we aim to solve two critical problems in previous image matching-based correspondence learning methods: 1) fail to fully exploit face specific structure information in correspondence estimation and 2) fail to learn personalized correspondence for each probe image. To this end, we first build a model, termed as morphable displacement field (MDF), to encode face specific structure information of semantic correspondence from a set of real samples of correspondences calculated from 3D face models. Then, we propose a maximal likelihood correspondence estimation (MLCE) method to learn personalized correspondence based on maximal likelihood frontal face assumption. After obtaining the semantic correspondence encoded in the learned displacement, we can synthesize virtual frontal images of the profile faces for subsequent recognition. Using linear discriminant analysis method with pixel-intensity features, state-of-the-art performance is achieved on three multipose benchmarks, i.e., CMU-PIE, FERET, and MultiPIE databases. Owe to the rational MDF regularization and the usage of novel maximal likelihood objective, the proposed MLCE method can reliably learn correspondence between faces in different poses even in complex wild environment, i.e., labeled face in the wild database.**

***Index Terms*— Face recognition, pose-invariant face recognition, 3D face model, 2D displacement field.**

I. Introduction

AFTER decades of research, many face recognition systems have been able to accurately recognize faces under controlled imaging conditions, as shown in the latest large scale evaluation [1]. Since the superiority of template matching method over geometric feature based method is discovered in the landmark work [2], most face recognition technologies represent faces as high dimensional vectors and compute their similarities according to some distance metrics defined in the vector space [3]–[6]. Consequently, most face recognition approaches rely on accurate face alignment to ensure the semantic correspondences between the features to match. Typically, the face alignment, i.e. spatial registration among faces, is accomplished by similarity or affine transformations in 2D image space, based on some manually or automatically located facial landmarks (e.g. eye centers).

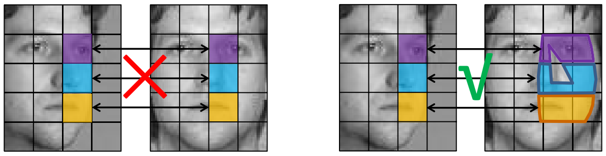


Fig. 1. Loss of semantic correspondence between faces under different poses using vectorized image representation.

However, in many real-world environments, large variations in pose will appear in face images. In this case, semantic correspondence between features cannot be easily reached by simple similarity or affine transformation, due to the complexity of human head in 3D structure. Therefore, most existing face recognition methods will fail in scenarios with faces presented in different poses. As shown in recent Multiple-Biometric Evaluation [1], the error rates of all the systems participating the test are significantly increased when dealing with non-frontal face images. So, handling large variation in pose is still one of the greatest technical challenges in face recognition field.

The main difficulty in Face Recognition Across Pose (FRAP) is the loss of semantic correspondence between feature dimensions, which is generally caused by the semantic misalignment between face images to compare. As exemplified in Fig. 1, given two face images of different poses aligned by affine transformation based on the eye centers, the semantics of a pixel (or an image block) in one image can be very different from that of the pixel (or image block) at the same spatial position in the other image. Please note that here by we term "semantics" as the biological meaning of the surface point(s) that a pixel (image block) represents.

Semantic misalignment is intrinsically caused by non-planar geometrical structure of human head, therefore this problem can be ideally solved if we have the 3D model of the

Manuscript received December 13, 2013; revised April 24, 2014; accepted July 18, 2014. Date of publication August 22, 2014; date of current version September 12, 2014. This work was supported by the Natural Science Foundation of China under Contract 61390511, Contract 61025010, Contract 61173065, Contract 61222211, and Contract 61001193. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Gang Hua. *(Corresponding author: Shiguang Shan.)*

S. Li, X. Liu, X. Chai, and S. Shan are with the Key Laboratory of Intelligent Information Processing, Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China (e-mail: shaoxin.li@vipl.ict.ac.cn; xin.liu@vipl.ict.ac.cn; xjchai@ict.ac.cn; sgshan@ict.ac.cn).

H. Zhang and S. Lao are with the Core Technology Center, OMRON Corporation, Kyoto 619-0283, Japan (e-mail: haihongzhang@oss.omron.co.jp; lao@ari.ncl.omron.co.jp).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TIP.2014.2351265

input faces. Based on this idea, Blanz and Vetter [7] proposed 3D Morphable Model (3DMM), which can fit a statistical 3D face model to one input 2D face image. Although 3DMM method can achieve very high accuracy for FRAP, it heavily relies on accurate 3D reconstruction from single image, which is however another big challenge. Additionally, due to its requirement to learn the statistical model from a 3D face database (containing faces with both 3D shape and texture), it is unclear how well it can generalize to novel face with unknown shape and texture [8].

To avoid the challenging 3D face recover problem, image matching based approaches are attracting more and more attention recently. Many methods [9]–[12] have shown promising capability of learning plausible semantic correspondence without reconstructing the 3D face model. In these methods, the semantic correspondence between two face images can be found by minimizing the sum of squared intensity difference between corresponding facial pixels, which actually assumes that the pixels of the same semantics should have similar intensities, especially when the two images are from the same subject. Formally the semantic correspondence between two face images can be expressed as a 2D Dense Displacement Fields (DDF), containing the displacement vectors linking the pixels with the same semantics in the two faces to match. As only 2D DDF rather than 3D shape and texture is needed to be learned, the parameters of image matching task is significantly reduced and the generalization ability is improved accordingly.

However, these image matching based methods also suffer from two problems when applying to FRAP problem. First, they fail to build an explicit prior model of plausible DDFs to regularize the displacement vectors between two images of the same face under varying poses. In previous image matching works, many priors are used, such as smoothness [11], slant assumption [12], cylinder model [9], ellipsoid model [13]. However, these priors are too general to characterize the special structure of DDF between faces under different poses. Second, these methods fail to learn personalized correspondence for each testing image. Ideally, in case that all the gallery images are in frontal pose, we only need to learn the correspondence between the probe face image and its frontal counterpart, i.e. personalized correspondence. However, previous image matching based methods either learn single correspondence for different probe images in the same non-frontal pose [9], [10], [13] or learn correspondence for arbitrary pair of gallery-probe images [11], [12].

To address the first problem, this paper follows the paradigm of image matching for face recognition across poses, but we propose to leverage the priors of 2D correspondence between face images of different poses learned from a 3D face database. Specifically, we first generate many ground truth correspondences from 3D faces. We call these correspondences as Template Displacement Fields (TDFs), from which a statistical representation model, Morphable Displacement Field (MDF), is built to constrain plausible correspondence as a convex combination of the TDFs. As the TDFs are generated from 3D faces, our method implicitly makes use of the 3D structural information of the faces.

To address the second problem aforementioned, we further propose an image specific correspondence learning method, named Maximal Likelihood Correspondence Estimating (MLCE), based on the MDF representation of correspondences between faces under different poses. The basic idea of the method is that a desirable DDF transformation should be able to generate a plausible virtual face image under the target pose (hereinafter, without losing generality, we assume the target pose be frontal). More specifically, we propose to estimate the image specific correspondence (i.e., DDF, expressed by our MDF) via maximizing the likelihood of the virtual frontal face image generated with the estimated correspondence. Finally, with the resulting correspondence, we can synthesize a virtual frontal image for subsequent recognition using classic face recognition methods such as Fisherfaces [3].

The schema of our MLCE method for face recognition across poses is shown in Fig. 2. Overall speaking, our method has two steps. In the first step, as shown in Fig. 2(a), given a probe non-frontal face image, its most plausible frontal face image is generated via image transformation by using the optimized DDF represented with MDF model. The optimization is targeted to maximize the likelihood of the virtual frontal face image generated by the DDF, in the Probabilistic PCA space learned from a reference frontal face image set. The second step, as shown in Fig. 2(b), is a standard frontal face recognition process, only with the difference that a self-occlusion mask estimated from the DDF, is applied to all the frontal face images in the gallery to adapt the virtual frontal face image. In this work, Fisherfaces with nearest neighbor classifier is exploited as the face recognition method.

This is a extended version of previous work [14]. Compared with the conference version, this paper has made three major extensions: 1) An MLCE method to estimate correspondence based on maximal likelihood criterion is proposed in Section IV-B; 2) More detailed comparisons and discussions are provided to analyze different criterions for learning optimal DDF in Section VI-B and 3) More comprehensive evaluations are conducted to validate MLCE method in Section VI-E.

The rest of the paper is organized as follows: Section II briefly reviews the previous works for face recognition across poses. Section III defines DDF and its representation via MDF model. We then present in Section IV the Maximal Likelihood Correspondence Estimation method. The next section describes how to conduct Face Recognition Across Pose with synthesized virtual frontal image, followed by comprehensive experimental evaluations in Section VI. Finally, we conclude the work and discuss possible future efforts in the last Section.

II. Related Works

As aforementioned, the grand challenge of Face Recognition Across Pose (FRAP) is intrinsically caused by terrible misalignment. Therefore, any method that addresses FRAP problem must have appropriate mechanism to build the semantic correspondence between faces in different poses. From this perspective, we classify previous FRAP algorithms into two categories: 1) Implicit Correspondence Learning (ICL);

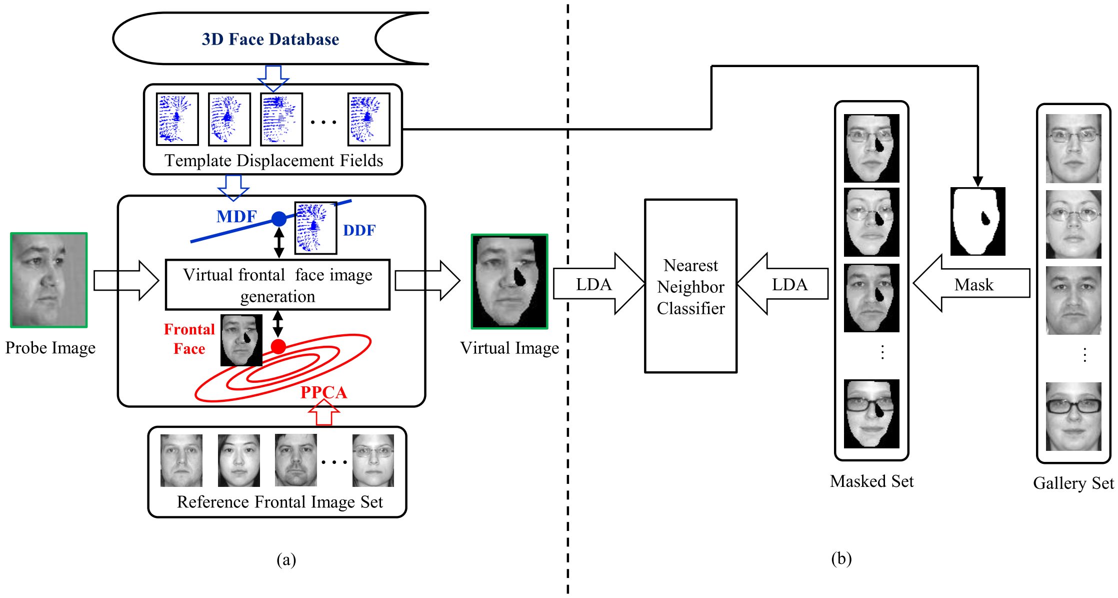

3D Face Database
Template Displacement Fields
MDF
DDF
Virtual frontal face image generation
Frontal Face
PPCA
Probe Image
Virtual Image
Reference Frontal Image Set
(a)
LDA
Nearest Neighbor Classifier
LDA
Mask
Masked Set
Gallery Set
(b)

Fig. 2. Schema of the proposed Maximal Likelihood Correspondence Estimation (MLCE) method for Face Recognition Across Pose (FRAP). (a) Maximal Likelihood Correspondence Estimation (MLCE) for image specific correspondence learning. (b) Face recognition based on synthesized virtual frontal face image with Linear Discriminant Analysis (LDA) method.

TABLE I
CATEGORIZATION OF METHODS FOR FRAP IN A CORRESPONDENCE LEARNING PERSPECTIVE

ICL	Common Space Learning [16], [17], [18]
	Pose Invariant Feature Extraction [19], [20], [21], [22]
ECL	Pose Specific [23], [9], [10], [24]
	Image Specific [7], [25], [26]
	Image Pair Specific [11], [12], [14]

2) Explicit Correspondence Learning (ECL). In each category, there are several distinctive subclasses, as summarized in Table I. In this section, we briefly review them to better position this work. For a more thorough survey, the readers are referred to [15].

A. Implicit Correspondence Learning

As its name suggests, methods in the "Implicit Correspondence Learning" category do not have an explicit procedure of dense correspondence building between face images under different poses. Instead, they only align the faces very coarsely by using very few facial landmarks (e.g. eye centers) as anchor points. Then, to facilitate the subsequent classification, they extract some features that are insensitive or immune to semantic misalignment. Such features can be obtained either by automatically learning a common space for different poses in a pure data-driven manner or manually designing some predefined pose-invariant measurements. Accordingly, we further classify the "Implicit Correspondence Learning" category of methods into two subcategories: 1. Common space learning; 2. Pose-invariant feature extraction.

Face images of different poses lie in different subspaces of the whole image space and thus cannot be compared directly, when they are only aligned coarsely. Common space learning methods attempt to seek for a common space where all the face images of varying poses are comparable. Prince et al. [16] exploit factor analysis model to construct a pose-invariant "identity subspace" for recognition. In [17], Sharma et al. propose to use Partial Least Square and Canonical Correlation Analysis to learn such common space. In [18], Kan et al. propose a multi-view discriminant analysis method to directly learn a common discriminant subspace for varying poses, in which Fishers separability criterion is maximized.

Approaches in the second subcategory attempt to define and extract some features that can be preserved even when the pose of the face changes. For instance, some methods first find a pose invariant statistics that can be calculated from full 3D face model. Then, after extracting incomplete feature of the statistics from a single input probe image, these approaches propose to recover complete statistics for final recognition. As the invariance of the feature is always ensured by some kinds of correspondence, this subclass of methods actually uses some prior correspondence information implicitly. One of such statistics is defined based on the linear object class assumption [27], which suggests that the combination weights of 3D face model are invariant to pose differences. In [20], Li et al. attempt to learn 3D model combination weights from 2D images by coupled bias-variance tradeoff regression. Build on the similar assumption that the nearest neighbor of a face are shared across different poses, in [21], Yin et al. propose an Associate-Predict model to associate 2D images under one pose to predict the appearance under another pose for recognition.

B. Explicit Correspondence Learning

By "explicitly correspondence learning," we mean that there are explicit steps which use spatial correspondence prior or learn such spatial correspondence for face recognition across pose. In this category, methods make a tradeoff between the correspondence precision and model complexity. Based on the precision of the correspondence, we further divide methods in this category into three subclasses: 1. Pose specific

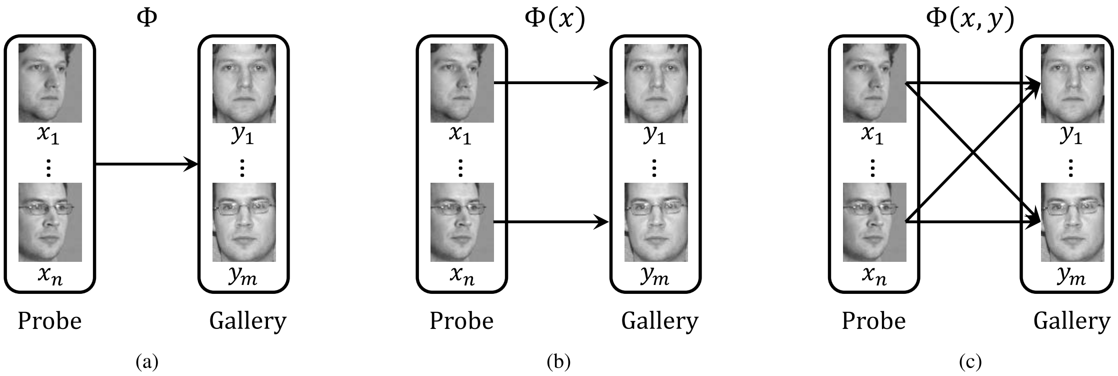


Fig. 3. Three categories of approaches learning explicit correspondence between face images in different poses: (a) Pose specific correspondence learning, i.e., all the faces under one non-frontal pose share the same correspondence field (to their frontal images); (b) Image specific correspondence learning, i.e., each non-frontal face image has its own correspondence to its own frontal face image; (c) Image-pair specific correspondence learning, i.e., the correspondence between each pair of images is needed to learn.

correspondence learning, which shares correspondence among faces under the same pose, shown in Fig. 3(a); 2. Image specific correspondence learning, which learns correspondence for each image, shown in Fig. 3(b); 3. Image pair specific correspondence learning, which learns correspondence for each pair of images, shown in Fig. 3(c).

In the first subclass, correspondence used to handle pose variation is shared among all individual faces under the same view. Such rigid correspondence can be directly calculated from general 3D face model or learned from data with some geometric prior. In [23], Gao et al. use a cylindrical shape model to approximate 3D face shape model. In [9], Chai et al. also use a general 3D cylinder face model to ensure approximate patch level semantic correspondence, but propose a local linear regression method to learn more detailed pixel level appearance variation in each patch. Li et al. [24] use a mean 3D face shape model calculated from a 3D face database to obtain more accurate semantic correspondence. Ashraf et al. [10] propose to learn patch level correspondence using an image matching based formulation in a pure data driven manner. Asthana et al. [28] present a fully automatic FRAP system. After fitting a View-based AAM, the input face is projected onto the aligned mean 3D face shape model, which is then rotated to render a frontal view for FRAP using LGBP [29]. Although general correspondence is not flexible enough for all individual faces, as shown in [9], [10], [23], [24], and [28], with only such approximate correspondence, better FRAP performance can be reached.

The second subclass of methods propose to fit personalized shape model for each face image, the learned correspondence of which can be much more precise compared to that learned by the first subclass of methods. Blanz and Vetter [7] propose a 3D Morphable Model (3DMM) to fit a full 3D face model to one input 2D face image based on a statistical model of 3D human face samples. Then the learned 3D model is either used to deduce semantic correspondence for virtual frontal image synthesis or directly used as invariant feature for subsequent recognition. Jiang et al. [26] propose to estimate personalized 3D shape model from a set of 2D facial landmarks. Then they can synthesize face images in variant pose, illumination and expression for recognition. As faces coming from different persons may have significantly different 3D face shapes, before recognizing the face, learning personalized correspondence for each image maybe a theoretically better way to deal with pose variation.

However, as fitting 3D face model from single image is very difficult and ill-posed, the third subclass of methods proposes to directly learn semantic correspondence in 2D image space between arbitrary pair of faces. The main concern in methods of this class is to impose rational constraint on the matching parameter. In [11], Arashloo and Kittler impose local smooth constraint on the learned matching parameter and propose an MRF based method to find semantic correspondence. Castillo et al. [12] employed dynamic programming-based stereo matching algorithm to find correspondences between frontal and non-frontal faces. Their basic assumption is that the misalignment caused by pose variation only presents in horizontal direction, i.e. slant assumption. Our previous work [14] also fall in this subclass. We use a set of template displacement field to build a Morphable Displacement Filed (MDF) model to constrain correspondence learning solution.

Generally speaking, compared to pure data-driven methods, approaches implicitly or explicitly using shape prior to help handling pose variation may achieve better performance. It is important to note that, as the projection matrix used in "Common Space Learning" methods and pose invariant feature used in "Pose Invariant Feature Extraction" methods are both shared among images under the same pose, all methods belong to "Implicit Correspondence Learning" category can also be regarded as pose specific correspondence learning approaches. While "Explicit Correspondence Learning" category of methods may be more capable in learning precise correspondence, the "Implicit Correspondence Learning" category of methods can implicitly learn correspondence while consider discriminant feature extraction for classification at the same time. So both categories have their own advantages. As a result, each subclass has some best performing FRAP methods.

III. Representation Model of Semantic Correspondence

In this section, we clarify the mathematic definition of semantic correspondence termed Dense Displacement

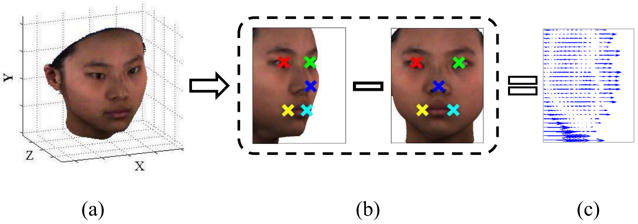


Fig. 4. Generation of template displacement field. (a) Original 3D face model; (b) Subtraction between pose specific normalized 3D face models; (c) Corresponding discrete 2D template displacement field. Note as only yaw pose is considered, most displacements in (c) are horizontal. But template displacement field is not limited to only deal with yaw pose.

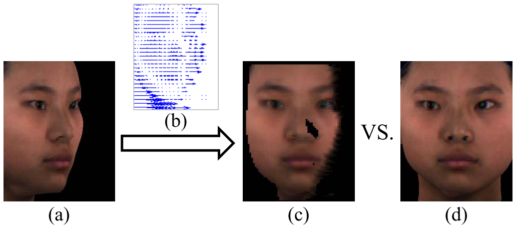


Fig. 5. Synthesized image using template displacement field. (a) Input profile face image; (b) Corresponding template displacement field; (c) Synthesized virtual frontal image; (d) True frontal face image.

Field (DDF) and validate the representation model of rational DDF termed Morphable Displacement Field (MDF). Given a 2D profile face image J_θ and its imaginary frontal counterpart J_0, the semantic correspondences can be described as a Dense Displacement Field (DDF) $\vec{d}_J$:

$$\forall t \in J_0 : J_0(t) = J_\theta(t + \vec{d}_J(t)). \tag{1}$$

Where $t = (x, y)$ is the coordinate of a pixel in 2D image J_0. For each pixel t in the frontal face image J_0, the corresponding pixel in the profile face J_θ locates in $t+\vec{d}_J(t)$, where each pixel of DDF $\vec{d}_J(t) = (\nabla x, \nabla y)$ indicates the displacement between them. As the 3D structures of faces are different, the DDF for different faces are not the same. This is the major obstacle in learning semantic correspondence. Fortunately, since different faces have similar structure, the DDFs must also have similar characteristics. In order to derive effective methods to learn DDF for different faces, we need to build a statistical model of DDF to make full use of such structure priors about human face at first.

A. Template Displacement Field

In previous works [11] and [12], DDF is forced to be smooth, which means that spatial relationship of neighborhood pixels in $J_\theta(t)$ changes steadily in the virtual image $J_\theta(t + \vec{d}_J(t))$. Such restriction is commonly used in computer vision tasks. However, it may be too weak for describing the structure of DDF between faces. Despite of such local properties, eligible DDF should also be globally consistent. To learn such global restriction of DDFs, we first calculate some prototype DDFs from 3D face shape models. The calculation procedure is sketched out in Fig. 4. Note, only shape model of 3D face model is used. We draw full 3D face model including texture in Fig. 4 only for visualization purpose.

Given a 3D face shape model, as shown in Fig. 4(a), one can rotate the model according to X or/and Y axis and project the rotated model onto X-Y plane to get 2D face shape model in arbitrary pose. Then with 2 or more landmarks, one can further normalize the face shape model in X-Y plane using Procrustes alignment method in order that the obtained template displacement fields are aligned in the same way of 2D images. Combining 3D rotation according to X or/and Y axes and 2D normalization in X-Y plane, we can generate pose specific 3D face shape models, see Fig. 4(b). Then the displacement caused by pose variation of each vertex on 3D face shape model can be calculated simply by coordinate subtraction between two pose specific normalized 3D face shape models. Finally, we use a 2D discrete displacement field to approximate the vertices' displacement in X-Y plane, and we term this discrete displacement field of realistic 3D face model as Template Displacement Field (TDF).

Ideally, with corresponding 2D TDF calculated from true 3D face model, a virtual frontal image can be synthesized from corresponding profile image simply by pixel interpolation as shown in Fig. 5. Due to facial region scaling in different poses, slight blur may be inevitably introduced in the synthesis procedure. But, in the experiments, we find such slight blur does not hinder classifiers from recognizing the face. It is also important to indicate that, as partial facial regions are occluded, only visible parts in profile view can be synthesized. As a result, the virtual image shown in Fig. 5(c) has some black undefined regions. Although such undefined regions of different faces in the same pose are slightly different, we only use rigid mask (details about the mask is illustrated in Section V) to remove most of the undefined regions in virtual image for recognition purpose.

B. Morphable Displacement Field

Although TDF of given probe image is not available, TDFs calculated from a reference 3D face database can help build more reasonable and valuable constraints on the feasible DDF. Therefore, after calculating a set of TDFs $\{\vec{d}_i \mid i = 1, 2, \ldots, N\}$, inspired by Blanz et al. [7], we propose to build a morphable representation model of DDF from them. The morphable face model presented in [7] is based on a vector space representation of faces that any convex combination of a set of aligned shape vectors S_i describes a realistic face shape vector S:

$$S = \sum_{i=1}^{N} \alpha_i S_i; \quad s.t. \sum_{i=1}^{N} \alpha_i = 1, \quad \alpha_i \geq 0. \tag{2}$$

Note the shape vector S and S_i are $V \times 3$ matrices which are consist of 3D coordinates of V vertices of the corresponding 3D face models. According to [27], assuming 3D face shape vectors approximately consist a linear object class. The coefficient α_i of Eqn. (2) will stay unchanged when linear operator L, e.g. 3D similarity transformation and 3D to 2D projection, is applied to shape model S:

$$S' = \sum_{i=1}^{N} \alpha_i S'_i; \quad s.t. \; S' = L \cdot S, \; S'_i = L \cdot S_i. \tag{3}$$

Since all the operators used in the calculation procedure of a TDF $\vec{d}_i$ are linear, the overall process shown in Fig. 4 can be formulated as:

$$\vec{d}_i \approx L_0 \cdot S_i - L_\theta \cdot S_i, \tag{4}$$

where S_i is a 3D face shape vector. L_0 and L_θ are linear operators applied to S_i in order to get pose specific normalized 3D face models shown in Fig. 4(b). Note, dense vertex-to-vertex alignment of 3D face shape model is done before we calculate TDF. Thus L_0 and L_θ corresponding to different 3D face shape models are almost the same. That is to say the linear operator applied to 3D face shape model is insensitive to the specific geometry of the face. Consequently, by combining Eqn. (2), (3) and (4), we can approximately express a realistic displacement field between a new pair of face images as the convex combination of pre-prepared TDFs:

$$\begin{aligned}\vec{d} &= L_0 \cdot S - L_\theta \cdot S \\ &= \sum_{i=1}^{N} \alpha_i \left[L_0 \cdot S_i - L_\theta \cdot S_i \right] \\ &\approx \sum_{i=1}^{N} \alpha_i \vec{d}_i. \end{aligned} \tag{5}$$

We term this formulation of realistic DDF as Morphable Displacement Field (MDF) model. Continuous changes of the model parameters α_i will generate smooth transition of target DDF $\vec{d}$ and any DDF $\vec{d}$ that can be represented as expression (5) must come from a realistic face.

IV. Correspondence Estimation for FRAP

After building MDF model to represent rational DDF, in this section, we present how to estimate correspondence between faces in different poses for subsequent FRAP. As shown in expression (1), the semantic correspondence of a profile face J_θ and its frontal counterpart J_0 can be represented as a DDF. Ideally, the rationality of such DDF $\vec{d}$ can be evaluated by the sum of squared intensity differences between indicated corresponding pixels in J_θ and J_0. After plugging MDF formulation of $\vec{d}$ shown in expression (5), the objective of MDF based correspondence estimation can be formulated as:

$$\begin{gathered}\alpha^* = \arg\min_{\alpha_i} \parallel J_0(t) - J_\theta(t + \sum_{i=1}^{N} \alpha_i \vec{d}_i(t)) \parallel_2, \\ s.t. \sum_{i=1}^{N} \alpha_i = 1, \quad \alpha_i \geq 0. \end{gathered} \tag{6}$$

However, in face recognition scenario, the identity of probe profile face is not known. As a result, frontal counterpart of probe profile face cannot be directly used as matching target to estimate DDF for FRAP.

A. Image Matching for FRAP

Typically, image matching based FRAP methods adapt expression (6) for face recognition by traversing all gallery frontal image $I_0(t)$ assuming that the best match will be reached when images I_0 and J_θ come from the same face. In other words, it proposes to minimize the matching residual between virtual frontal image of probe face and each gallery frontal image:

$$\begin{gathered}\alpha^* = \arg\min_{\alpha_i} \parallel I_0(t) - J_\theta(t + \sum_{i=1}^{N} \alpha_i \vec{d}_i(t)) \parallel_2, \\ s.t. \sum_{i=1}^{N} \alpha_i = 1, \quad \alpha_i \geq 0. \end{gathered} \tag{7}$$

B. Probabilistic Matching for FRAP

Although traversal scheme can be used to learn the sematic correspondence between faces in different poses for face recognition, this scheme has two problems: 1. Would be time consuming — correspondence needs to be learned with every gallery image; 2. May suffer from over-fitting — even the face images I_0 and J_θ come from different persons, the correspondence is learned to match them.

To solve these problems, a new probabilistic matching method based on maximal likelihood criterion is proposed to estimate the correspondence. Intuitively, before recognizing $J_\theta(t)$, we hope the synthesized image $J_\theta(t + \vec{d}(t))$ to be a rational frontal face as likely as possible. Since the intensity of synthesis image $J_\theta(t + \vec{d}(t))$ comes from probe image $J_\theta(t)$ and rational DDF would not cause undesirable distortion that changes the identity of probe image, the synthesis image $J_\theta(t + \vec{d}(t))$ can be a maximal likelihood frontal face only when it is similar to the true frontal view of probe image. Assuming the frontal face images distribute as a Gaussian, the objective of probabilistic matching can be formulated as:

$$\begin{gathered}\alpha^* = \arg\max_{\alpha} P(J_\theta(t + \sum_{i=1}^{N} \alpha_i \vec{d}_i(t)) | \Omega_0), \\ s.t. \sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0, \end{gathered} \tag{8}$$

where $\Omega_0 = (\mu_0, \Sigma_0)$ is the parameter of the Gaussian distribution model of general frontal faces which can be used to evaluate the probability of an image to be a rational frontal face. As synthesis image is not forced to match any specific image, this probabilistic matching is not likely to suffer from over-fitting.

In this paper, we use Probabilistic PCA [30], which is a special case of common Factor Analysis, to build the probabilistic model of general frontal faces. Note, only frontal images are used to build Probabilistic PCA model. Thus we remove subscript "$*_0$" for frontal image or related parameters to make it pithy. According to typical FA model, we first formulate likelihood of frontal face image vector $x \in \mathbb{R}^M$ as a Gaussian, the mean of which is a linear function of hidden variable $z \in \mathbb{R}^m$:

$$P(x|z, \Gamma) = \mathcal{N}(Wz + \mu, \Psi), \tag{9}$$

where $\Gamma = (W, \mu, \Psi)$. Commonly the distribution of z is also assumed to be a Gaussian and can be set to be standard

Gaussian $\mathcal{N}(\mathbf{0}, I^{m\times m})$ without loss of generality. More specifically, in Probabilistic PCA, the covariance matrix Ψ is assumed to be isotropic, in other words $\Psi = \sigma^2 I^{M\times M}$. Linear transformation matrix $W \in \mathbb{R}^{D\times L}$ is assumed to be orthogonal. Therefore, the z-conditional probability distribution over x-space can be further expressed as:

$$\begin{aligned} P(x|z,\Gamma) &= \mathcal{N}(Wz+\mu, \sigma^2 I^{M\times M}),\\ P(z) &= \mathcal{N}(\mathbf{0}, I^{m\times m}). \end{aligned} \tag{10}$$

With expression (10) the marginal distribution for the observed image x is readily obtained by integrating out the latent variables and is likewise Gaussian:

$$\begin{aligned} P(x|\Gamma) &= \int P(x|z,\Gamma)P(z)dz,\\ &= \mathcal{N}(\mu, WW^T + \sigma^2 I^{M\times M}) \end{aligned} \tag{11}$$

Given a set of training frontal face images $X \in \mathbb{R}^{M\times N}$, the covariance matrix of which can be decomposed as $XX^T = V\Lambda V^T$, then the maximal likelihood estimation of parameter $\Gamma = (W, \mu, \sigma)$ of Probabilistic PCA model is:

$$\begin{aligned} \hat{\mu} &= \frac{1}{N} X \cdot \mathbf{1}^{N\times 1}, \quad \hat{\sigma}^2 = \frac{1}{M-m}\sum_{j=m+1}^{M} \lambda_j^2,\\ \hat{W} &= V_m(\Lambda_m - \hat{\sigma}^2 I^{m\times m})^{\frac{1}{2}}. \end{aligned} \tag{12}$$

Note that Λ_m is the diagonal matrix of the largest m eigenvalues of Λ and V_m is consist of the corresponding columns of V. Once we obtain the parameter Γ of Probabilistic PCA model, we can calculate the probability of a image to be a frontal face with expression (11). The objective shown in expression (8) can now be rewritten as:

$$\begin{aligned} \alpha^* &= \arg\min_{\alpha} \parallel \mu_0(t) - J_\theta(t + \sum_{i=1}^{N} \alpha_i \vec{d}_i(t)) \parallel_{\Sigma_0},\\ &s.t. \sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0. \end{aligned} \tag{13}$$

where $\mu_0 = \hat{\mu}$ is the mean frontal face and $\parallel \cdot \parallel_{\Sigma_0}$ denotes Mahalanobis distance with weight matrix Σ_0 and $\Sigma_0 = (\hat{W}\hat{W}^T + \hat{\sigma}^2 I)^{-1}$. Note this is very similar to image matching by means of the Mahalanobis distance, except that only mean frontal face is used as matching target. In practice, we find expression (13) performs fairly well in learning personalized correspondence without blindly traversing all gallery images. As a result, probabilistic matching not only reduces the potential risk of over-fitting, but also significantly accelerates the whole recognition procedure. As the correspondence estimated with expression (13) can generate maximal likelihood frontal face from probe non-frontal image J_θ, we term this method as Maximal Likelihood Correspondence Estimation (MLCE).

C. Optimization of MLCE

In practice, we find typical gradient based methods [31] and [32] fail to obtain satisfactory solution for MDF based correspondence estimation, i.e. Eqn. (7) and Eqn. (13). As a result, we need a more effective way to optimize MDF.

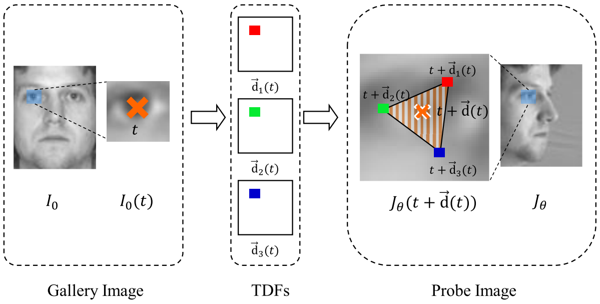


Fig. 6. Example of image matching procedure indicated by Eqn. (7) when there are 3 template displacement fields.

Before finding such an optimization method, we first investigate how MDF works fundamentally. As shown in Fig. 6, each pixel t of a TDF $\vec{d}_i$ indicates a candidate matching point $t + \vec{d}_i(t)$ in probe image J_θ of pixel t in gallery image I_0. Since DDF $\vec{d}(t)$ is a convex combination of TDFs $\vec{d}_i$, in geometry, feasible region of true matching point is a convex hull of N vertices determined by N corresponding TDFs. For example, when $N = 3$, for each pixel t in image I_0, the feasible region of matching point is actually a triangle, as shown in Fig. 6. Considering that, if the feasible region is sufficiently small, then gray-scale intensity of any pixel in it can be approximately interpolated by gray-scale intensity of the convex hulls vertices:

$$J_\theta(t + \sum_{i=1}^{N} \alpha_i \vec{d}_i(t)) \approx \sum_{i=1}^{N} \alpha_i J_\theta(t + \vec{d}_i(t)). \tag{14}$$

Actually using simple calculus it can be proved that the first order Taylor Expansion of the left and right sides of Eqn. (14) are the same. With the approximation indicated in Eqn. (14), we can relax Eqn. (7) and (13) as:

$$\textbf{IM:} \quad \arg\min_{\alpha} \parallel I_0(t) - \sum_{i=1}^{N} \alpha_i J_\theta(t + \vec{d}_i(t)) \parallel_2, \tag{15}$$

$$\textbf{PM:} \quad \arg\min_{\alpha} \parallel \mu_0(t) - \sum_{i=1}^{N} \alpha_i J_\theta(t + \vec{d}_i(t)) \parallel_{\Sigma_0}. \tag{16}$$

Note that "**IM**" represents Image Matching and "**PM**" represents Probabilistic Matching. The convex combination restriction is still used, but we omit it for simplicity. The optimization problem of Eqn. (15) and (16) are a quadratic programming. Since the objective is convex, the optimization has unique global minimum and can be effectively solved with common convex optimization methods in polynomial time. Compared to Eqn. (7) and (13), rationality of Eqn. (15) and (16) depends on the precision of approximation indicated by Eqn. (14). Therefore the size of the convex hull (see Fig. 6) should be controlled. To control the size of the convex hull, we first prune $J_\theta(t + \vec{d}_i(t))$ before the optimization and wipe out raw synthesis images which are far away from $I_0(t)$. On the other hand, as optimization of Eqn. (15) and (16) are actually a common regression problem with L1 regularization adding a non-negative constraint, many coefficients of optimal solution of Eqn. (15) and (16) will be further shrunk to zero.

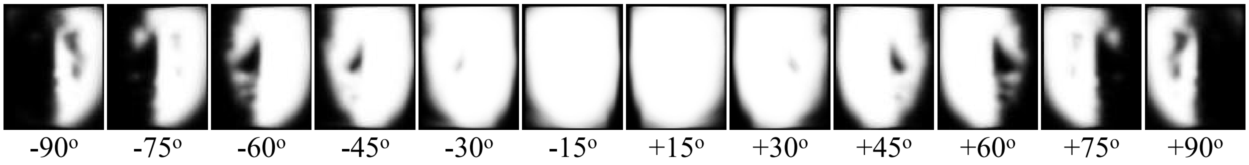


Fig. 7. Visible probability mask for different poses. The intensity of the mask image represents the probability about whether the pixel can be seen in a given pose.

Consequently, the preserved synthesis images $J_\theta(t + \vec{d}_i(t))$ with non-zero coefficients will be similar to gallery image, which indicates that the corresponding template displacement fields $\vec{d}_i(t)$ are similar. Thus the corresponding convex hull shown in Fig. 6 will be relatively small, which leads objective of Eqn. (15) and (16) be sufficiently close to original objective, i.e., Eqn. (7) and (13). What's more, with pruning step, the complexity of the optimization is also decreased. Finally, we obtain a more compact and rational version of Eqn. (15) and (16), which can be formulated as:

$$\textbf{IM:} \quad \arg\min_{\alpha_N} \| I_0(t) - \sum_{i=1}^{K} \alpha_{N_i} J_\theta(t + \vec{d}_{N_i}(t)) \|_2, \tag{17}$$

$$\textbf{PM:} \quad \arg\max_{\alpha_N} \| \mu_0(t) - \sum_{i=1}^{K} \alpha_{N_i} J_\theta(t+\vec{d}_{N_i}(t)) \|_{\Sigma_0}, \tag{18}$$

where $I_0(t)$ is gallery frontal image, $J_\theta(t + \vec{d}_{N_i}(t))$ is the ith nearest neighbor of $I_0(t)$ in all synthesized images. In the final objective in expression (17) and (18), we implicitly obtain displacement field $\vec{d}(t)$ between gallery and probe image. So we call it implicit Morphable Displacement Field (iMDF). In practice, for probabilistic matching K can be very small. By default, we set $K = 5$. However, we find even when $K = 1$ the recognition accuracy is acceptable. This may be because the 3D shapes of different faces are very similar, so that any test image can find a TDF to approximate its own DDF accurately.

V. Recognition

Once the virtual frontal image is synthesized, we can use common template matching based methods to recognize it. One of the virtual images synthesized by proposed probabilistic matching is shown in Fig. 2. Note the black region of virtual image is generated by visibility mask. Since part of profile face is not visible in frontal view, we generate mask from 3D face model in different poses to approximate the visible probability, as shown in Fig. 7. To calculate the probability, we first rotate all 3D face models to the target pose. Then, for each pixel, if it can be seen in k out of N 3D face models the visible probability is set as $\frac{k}{N}$.

In the training step, for each profile image, we pick K maximal likelihood virtual images $J_\theta(t + \vec{d}_{N_i}(t))$ together with the masked frontal image set (see Fig. 2) to train LDA classifiers. In the testing step, for each profile image, the virtual frontal image $\sum_{i=1}^{K} \alpha_{N_i} J_\theta(t + \vec{d}_{N_i}(t))$ obtained by minimizing expression (17) or (18) are used for recognition.

VI. Experiments

In this section, we systematically evaluate our method on three multi-pose face benchmarks, i.e., CMU-PIE [33], FERET [34] and MultiPIE [35] to verify its effectiveness. Although the DDF can be calculated between arbitrary pose pairs, in this paper, we only present results when gallery image has frontal pose and probe image has non-frontal pose. Images in this paper are normalized according to pose specific mean shape using Procrustes analysis. To obtain the mean shapes under different poses, we first manually labeled 5 landmarks (two eyes, nose tip and two mouth corners) on 700 3D face models in BJUT [36]. Then we rotated the 3D face models to different pose and used the corresponding coordinates in X-Y plane to calculate the mean shapes using generalized Procrustes alignment method. Note if the yaw angle is larger than 60°, one eye and one mouth corner will become invisible. So we only use 3 visible landmarks to align face when its yaw pose is larger than 60°. The mean shapes in different pose calculated from BJUT and examples of normalized face images in CMU-PIE, FERET and MultiPIE are shown in Fig. 8.

To ensure that normalized face images in different poses have similar size, we scale the pose specific mean shape in order that the distance between the horizontal line across eyes and mouth center (center point of two mouth corners) stays the same. As shown in Fig. 8, with this normalization method, all face images are normalized to 64 × 80. We also draw unified yaw pose rules above or below the sample images to show approximate yaw poses.

Totally, we conducted six sets of experiments and present the results in six corresponding subsections. At first, we compare different image matching based FRAP methods to shown the effectiveness of MDF model. Secondly, we compare four different objectives of MDF based correspondence learning to validate the rationality of the Probabilistic Matching criterion. Thirdly, we evaluate the robustness of probabilistic matching with pose estimation errors. After showing the influence of several parameters, i.e. preserved number of TDFs in pruning step and dimension of LDA classifier, we finally compare the proposed method with other state-of-the-art FRAP methods and general FR methods.

A. Comparison of Image Matching Based FRAP Methods

As aforementioned, correspondences between semantic facial points are highly structured. In order to obtain rational solution of image matching based correspondence learning, proper shape priors must be imposed on DDFs. In this section, we compare several image matching based methods which use different shape priors to learn correspondence.

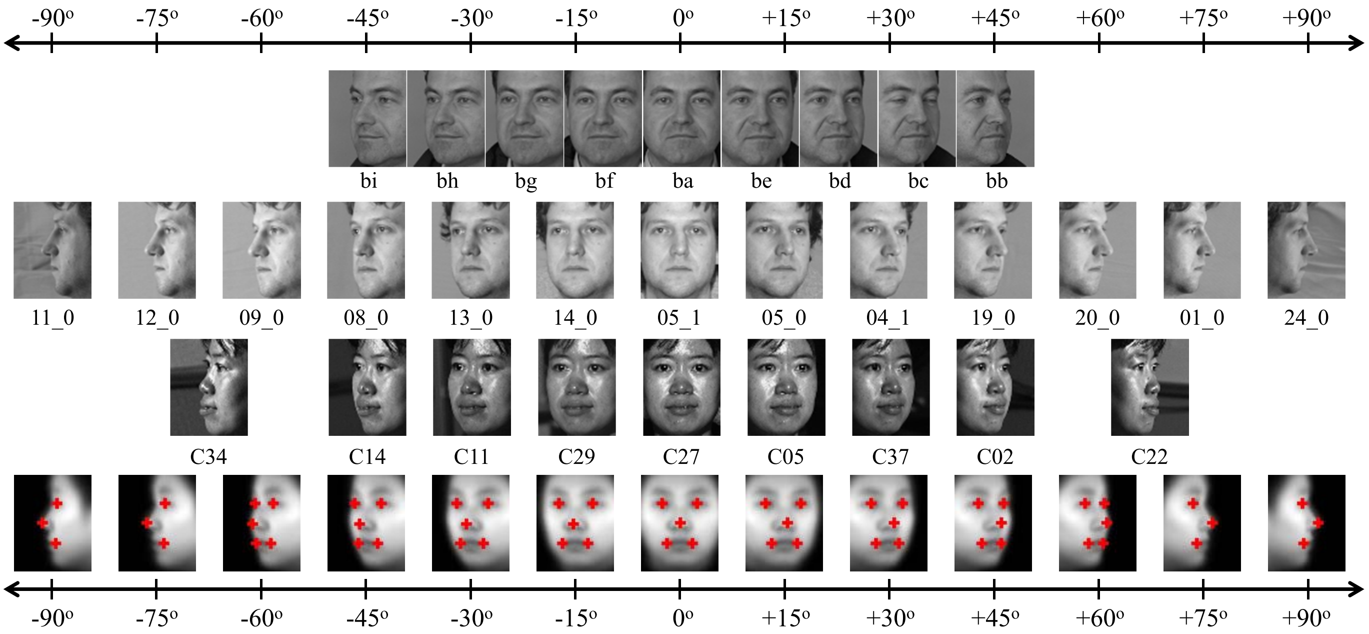
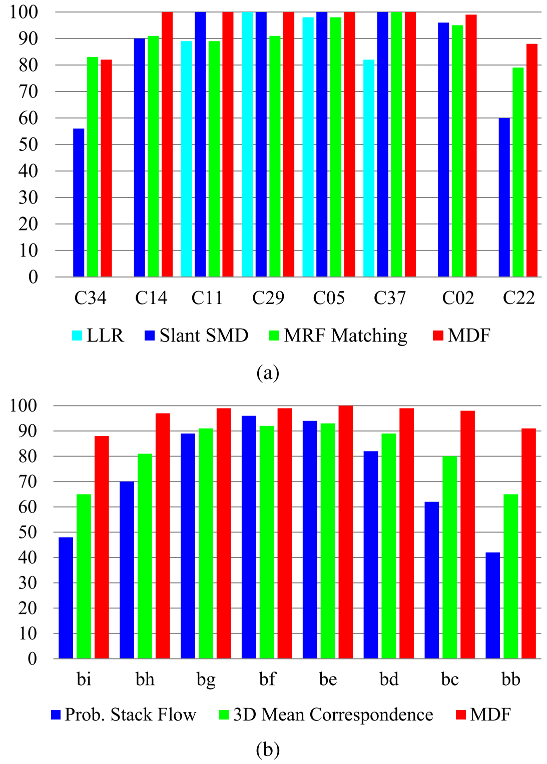


Fig. 8. Samples of normalized face images using pose specific alignment and the corresponding mean shapes of face in different poses. The samples in the first row come from FERET database. The samples in the second row come from MultiPIE database and the samples in the third row come from PIE database. The bottom row is the mean face generated from BJUT 3D face database and the corresponding mean shape of used facial landmarks. Above or below the sample images, there are unified rulers to indicate the approximate yaw angle of each image.



(a)

(b)

Fig. 9. Comparison of Different Image Matching Based FRAP methods on (a) CMU-PIE database and (b) FERET database.

Subsets of PIE database with only yaw pose and subsets $ba, bb, bc, bd, be, bf, bg, bh, bi$ of FERET database are used to conduct the experiments. The results are shown in Fig. 9. Note, except MDF based image matching, results of all other methods all obtained from the original papers.

Before analyzing the results, we summarize the comparison methods according to correspondence learning precision and strength of used shape prior. In the correspondence learning precision perspective (Please refer to Table I), Probabilistic Stack Flow [10], Local Linear Regression [9] and 3D Mean Correspondence [24] are pose specific correspondence learning method. Slant SMD [12], MRF Matching [11] and MDF base matching are image pair specific correspondence learning method. In the strength of used shape prior perspective, Probabilistic Stack Flow learns correspondence in a pure data-driven manner. Slant SMD and MRF Matching use general shape prior while Local Linear Regression, 3D Mean Correspondence and MDF use face specific shape prior.

In Fig. 9, we can see that learning image pair specific correspondence is better than learning pose specific correspondence. This validates the necessity of learning different correspondence for different faces due to its distinctive 3D structure. However, since image pair specific correspondence learning is easy to be over-fitting, generally speaking, using stronger priors is demonstrated to be beneficial to the final performance. For example directly using 3D Mean Correspondence is better than learning the correspondence in a pure data driven way as in Prob. Stack Flow and using statistical MDF model built from template DDF to regularize the solution is better than imposing only generic shape constraints introduced in Slant SMD, MRF matching.

B. Evaluation of Different Objectives for MDF Optimization

As mentioned in section IV, the ideal MDF based correspondence learning criterion expression (6) cannot be directly used for face recognition. To cope with this problem, we propose two solutions: image matching objective and probabilistic matching objective. In order to clarify the influence

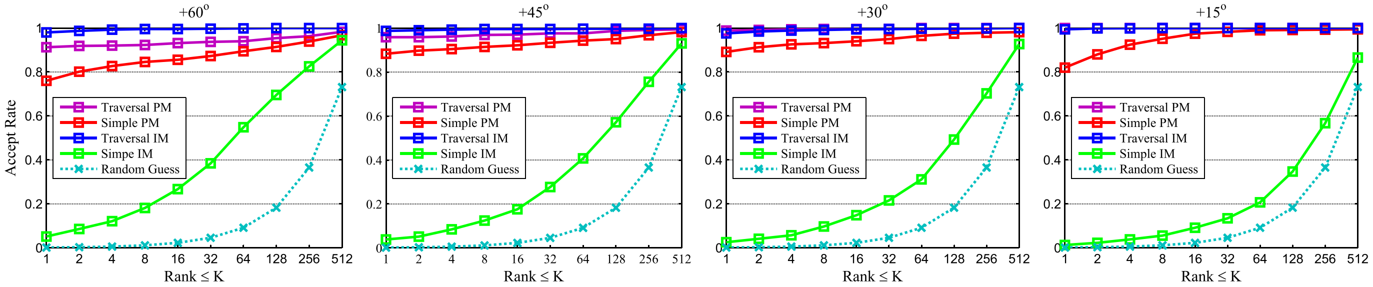


Fig. 10. Ground truth DDF ranking test on BJUT 3D database. As good objective should give highest preference for ground truth DDF, we rank TDFs according to values generated by objective functions and if the ground truth DDF is included in the top K DDFs, we accept the objective as a good evaluation of DDF. Note the cyan dotted line is the results of random guess.

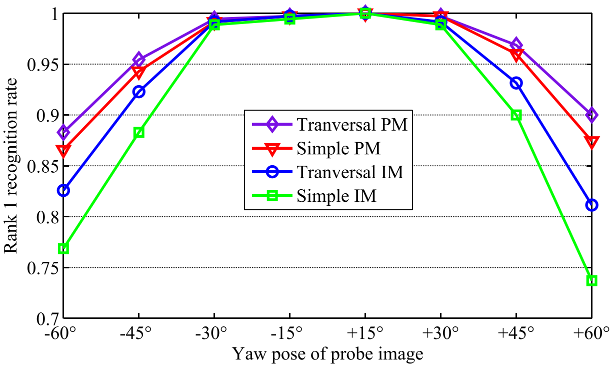


Fig. 11. Rank 1 recognition test on BJUT 3D database.

TABLE II
TERMINATION OF DIFFERENT OBJECTIVES FOR MDF OPTIMIZATION

Matching Method Name	Mathematical Formulation
Simple IM	$\parallel \mu_0(t) - \sum_{i=1}^{N} \alpha_i J(t + \overrightarrow{d_i}(t)) \parallel_2$
Traversal IM	$\parallel I_0(t) - \sum_{i=1}^{N} \alpha_i J(t + \overrightarrow{d_i}(t)) \parallel_2$
Simple PM	$\parallel \mu_0(t) - \sum_{i=1}^{N} \alpha_i J(t + \overrightarrow{d_i}(t)) \parallel_{\Sigma_0}$
Traversal PM	$\parallel I_0(t) - \sum_{i=1}^{N} \alpha_i J(t + \overrightarrow{d_i}(t)) \parallel_{\Sigma_0}$

of these adaptations, in this subsection, we compare these two objectives. Moreover, to further clarify the effects of the traversal strategy used in image matching objective and the Mahalanobis distance employed in probabilistic matching objective, we also present the results of using traversal strategy in the probabilistic matching and using simple mean frontal face image as matching target in image matching. The objectives of comparison methods are shown in Table II.

As there are 700 3D face models, we can generate 700 face images under any appointed pose. Then for each pose image, we can generate 700 virtual frontal images using TDFs. Intuitively, good objectives should give highest preference for ground truth DDF. Thus we first rank TDFs according to objective values. Then, if the ground truth DDF is included in the top K DDFs, we accept the objective as a good evaluation of DDF. As shown in Fig. 10, as the ground truth DDF between profile and frontal face images indeed exists in the 700 TDFs, the traversal IM and PM objective performs fairly well in such scenario. While the simple IM objective performs only slightly better than random guess, the simple PM objective without traversal strategy also achieves satisfiable performance.

Although experimental results show that the traversal strategy of image matching objective may give higher preference

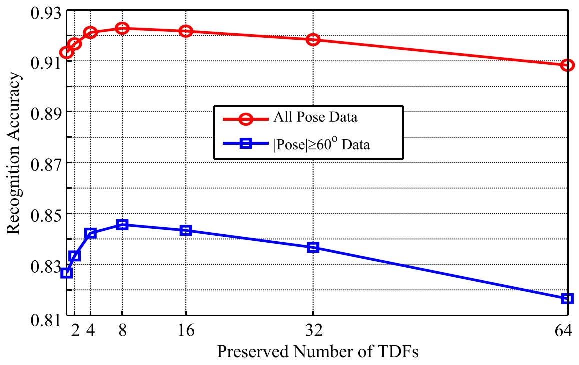


Fig. 12. Evaluation on sensitivity of preserved synthesized image number in the pruning step on MultiPIE database.

to ground truth DDF, in more practical condition, the ground truth will not be included in the TDFs. In order to evaluate objective functions in this circumstance, we divide the BJUT database into two equal parts and use one part's TDFs to build MDF model for DDF learning in another part. With nearest neighbor classifier based on Euclidean distance, the recognition rates are shown in Fig. 11.

It can be seen that both traversal and simple PM objectives outperform traversal IM objective in such situation. And the traversal PM performs even better than simple PM, this may indicate the necessity of traversal strategy when use regularized displacement field via MDF, which can help avoiding undesirable over-fitting and irrational matches between faces of different persons. However, it is also important to note that traversal strategy is much more time-consuming, because they need to traverse all gallery images for recognition purpose which leads to N (Gallery Size) times of computational cost. Thus the simple PM, which can achieve satisfiable performance with only $\frac{1}{N}$ time consumption of traversal PM, is much more suitable for real-time applications.

C. Evaluation of Parameter Sensitivity

One of the key for effective MDF based optimization is the approximation indicated by expression (14). As illustrated in section IV-C, we propose a pruning step to ensure the precision of the approximation. We present the mean recognition accuracy on MultiPIE database with different preserved number of TDFs after pruning step in Fig. 12. As shown in Fig. 12, when the preserved number of TDFs is too small or too large the mean recognition accuracy will decline. This is because

TABLE III

POSE ERROR TOLERANCE TEST OF MAXIMAL LIKELIHOOD DISPLACEMENT FIELD ON MULTIPIE DATABASE

Pose	11_0	12_0	09_0	08_0	13_0	14_0	Pose	05_0	04_1	19_0	20_0	01_0	24_0	Avg
Error	-75^o	-70^o	-60^o	-45^o	-30^o	-15^o	Error	$+15^o$	$+30^o$	$+45^o$	$+60^o$	$+70^o$	$+75^o$	
-15^o	61.7	83.2	97.3	99.3	100	100	$+15^o$	100	100	99.3	91.9	87.9	58.4	89.9
-10^o	64.4	89.3	96.6	100	100	100	$+10^o$	100	100	100	96.0	85.9	58.4	90.9
-5^o	65.8	91.3	96.6	100	100	100	$+5^o$	100	100	100	96.0	89.9	58.4	91.5
0^o	69.1	90.6	97.3	100	100	100	0^o	100	100	100	96.6	90.6	61.1	**92.1**
$+5^o$	61.7	90.6	98.0	99.3	100	100	-5^o	100	100	100	95.3	87.2	54.4	90.5
$+10^o$	57.0	87.2	96.6	98.7	100	100	-10^o	100	100	98.7	96.0	88.6	53.7	89.7
$+15^o$	53.7	85.2	94.0	97.3	100	100	-15^o	100	99.3	98.7	94.0	81.9	53.7	88.2
std	5.24	3.09	1.27	0.98	0	0	-	0	0.26	0.62	1.64	2.91	2.92	-

when too few TDFs are used, the built MDF model may be not flexible enough to represent variation of true DDF. When too many TDFs are used, the approximation indicted by expression (14) is not precise enough. We also find that when the pose angle is large, the MDF model with proper number of TDFs brings more considerable benefit.

D. Evaluation of Tolerance to Pose Estimation Error

Since DDF changes according to probe pose, given a probe face image captured in 45° yaw pose, it is best to use TDFs between 45° and 0° to build MDF model for learning correspondence of this probe image. However, the yaw pose cannot always be estimated accurately. Thus, it is important that our correspondence learning method can tolerate pose estimation errors to some extent.

In practice, the absolute pose estimation error may not exceed 15°. As shown in Table III, when the absolute error of pose estimation is less than 15°, the recognition accuracy will not decline considerably. If the yaw pose of probe image is small, the proposed probabilistic matching method is even more robust for pose estimation error. The last column of the table is the average recognition accuracy. Intuitively, if the pose of TDF is more close to the probe image, the recognition accuracy would be higher. In this paper, all the other results are obtained with the same pose setting that obtain best average recognition accuracy, as shown in Table III.

E. Comparison With State-of-the-Art FRAP Methods

We compare the proposed methods with state-of-the-art FRAP methods on three typical multi-pose databases, i.e. CMU-PIE, FERET and MultiPIE. To clarify different Train/Test divisions used in different methods, we mark different methods with different superscripts. In CMU-PIE database, superscript "$*^1$" indicate all data of 68 persons is used for testing while "$*^2$" means half of the data is used for training and the other data is used for testing. The LDA classifier of MDF-PM^1 is trained on MultiPIE database. In FERET database, the meaning of superscripts is similar. In MultiPIE database, there are three different kinds of Train/Test divisions used in previous works. The first division strategy only uses data captured in session 1 to conduct experiments. Data of first 100 persons is used for training and the remaining 149 persons' data is used for testing. We mark this division as "$*^1$." In the second kind of division marked as "$*^2$," data of all 4 sessions is used with 100 persons as training set and the remaining 237 persons as testing set. The training set and probe set of the third kind of division is the same as the second kind of division. But the gallery set of the third kind of division is forced to has only one image for each person. Such configuration would introduce more variation other than pure pose between the gallery and probe images. Thus, the third kind of division marked as "$*^3$" is the most challenging division. As shown in Table IV, using simple LDA classifier with gray-scale intensity feature, our methods outperform all the other approaches. In more challenging scenario, i.e. the third division in MultiPIE database, it seems that more robust classifier should be used to handle appearance variations between images in different sessions.

F. Comparison With State-of-the-Art General FR Methods

Finally, we compare MLCE with state-of-the-art face verification methods, i.e. Multiple One Shot Similarity (MOSS) [37], Probabilistic Elastic Matching (PEM) [38] and Fisher Vector (FV) [39]. These methods are able to handle many real world facial appearance variations, such as pose, expression and illumination, and achieve promising performance on wild environment face database, i.e. Labeled Face in the Wild (LFW) database [40]. To conduct fair comparisons, for all three method, dense SIFT feature extracted from 64×80 gray-scale image is used. Note, we implement MOSS and FV methods according to the original papers and achieve 83.68% and 86.2% mean accuracy on LFW respectively, which is comparable to the results reported in the original papers. For PEM method, we use the binary code provided by the author to conduct experiments and achieves 81.05% mean accuracy on LFW. The comparisons are conducted in two aspects: First, we present results of MOSS, PEM and FV on MultiPIE database in protocol 1 to evaluate their capability in handling pure pose variation. Second, we present results of PEM method with or without virtual frontal face synthesis to evaluate the effectiveness of MLCE in handling pose variation in real world face verification tasks. As shown in Table V, the PEM method which uses elastic matching to cope with pose variation performs better than MOSS and FV method that do not explicitly address pose problem. And the proposed MLCE method can achieve best FRAP performance with the LDA method using synthesized virtual frontal image as feature.

After validating the excellence of MLCE in handling pure pose variation, we further evaluate the capability of MLCE in handling pose variation under the so called "wild" environment. More specifically, we present results of PEM method on LFW with or without MLCE based virtual frontal

TABLE IV

FRAP PERFORMANCE COMPARISON WITH STATE-OF-THE-ART FRAP METHODS. (a) CMU-PIE. (b) FERET. (c) MULTIPIE

(a)

Methods	C34	C14	C11	C29	C05	C37	C02	C22	Avg	Avg
	-65^o	-45^o	-30^o	-15^o	15^o	30^o	45^o	65^o	$C11-C37$	$C34-C22$
Castillo11[12][1]	56	90	100	100	100	100	96	60	100	87.8
Arashloo11[11][1]	83	91	89	91	98	100	95	79	94.5	90.8
Li12[20]-Gabor[2]	72	87	100	100	100	100	100	74	100	91.6
Sharma12[17][2]	85	97	100	100	100	100	85	79	100	93.2
MDF-IM[1]	82	100	100	100	100	100	99	88	100	96.1
MDF-PM[1]	91	99	100	100	100	100	99	87	100	97.0
MDF-PM[2]	98.5	100	100	100	100	100	100	97.1	100	99.5

(b)

Methods	bi	bh	bg	bf	be	bd	bc	bb	Avg
	-65^o	-45^o	-30^o	-15^o	15^o	30^o	45^o	65^o	
Blanz03[7][1]	95	95	97	100	97	96	95	91	95.8
Li12[20]-Gabor[2]	78	91	96	96	98	99	96	87	92.6
Sharma12[17][2]	79	85	94	96	95	94	82	70	86.4
MDF-IM[2]	88	97	99	99	100	99	98	91	96.4
MDF-PM[2]	97	99	99	99	100	100	100	99	99.1

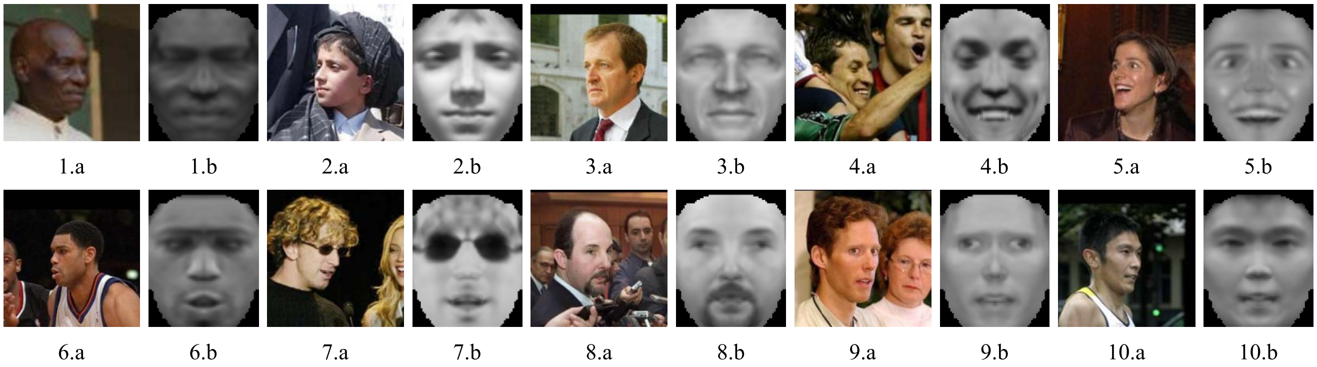
(c)

Methods	11_0	12_0	09_0	08_0	13_0	14_0	05_0	04_1	19_0	20_0	01_0	24_0	Avg
	-75^o	-70^o	-60^o	-45^o	-30^o	-15^o	$+15^o$	$+30^o$	$+45^o$	$+60^o$	$+70^o$	$+75^o$	
Li12[20][1]	42.3	60.4	76.5	92.0	97.3	100	100	98.0	85.9	75.8	60.4	41.6	77.5
Li12[20]-Gabor[1]	32.2	69.1	87.9	97.3	99.3	100	100	100	92.6	91.3	71.1	41.6	81.9
MDF-PM[1]	69.1	90.6	97.3	100	100	100	100	100	100	96.6	90.6	61.1	92.1
Sharma12[17][2]	27.0	42.2	48.5	84.8	96.6	99.2	99.2	96.2	89.0	57.4	47.7	27.8	68.0
MDF-PM[2]	53.2	80.0	89.8	95.7	99.0	100	100	99.3	95.5	88.3	77.2	45.5	85.3
Asthana11[28][3]	-	-	-	74.1	91.0	95.7	95.7	89.5	74.8	-	-	-	86.8
MDF-PM[3]	-	-	-	90.0	94.3	95.3	94.7	93.7	87.7	-	-	-	92.6

TABLE V

FRAP PERFORMANCE COMPARISON WITH GENERAL FACE RECOGNITION METHOD ON MULTIPIE

Methods	11_0	12_0	09_0	08_0	13_0	14_0	05_0	04_1	19_0	20_0	01_0	24_0	Avg
	-75^o	-70^o	-60^o	-45^o	-30^o	-15^o	$+15^o$	$+30^o$	$+45^o$	$+60^o$	$+70^o$	$+75^o$	
grayscale-LDA[1]	20.8	33.6	40.9	74.5	91.3	99.3	99.3	90.6	75.8	38.3	30.9	17.4	59.4
MOSS+ID+pose[37][1]	24.8	43.6	54.4	89.9	98.0	99.3	100	98.0	89.3	52.3	44.3	20.8	67.9
PEM[38][1]	18.8	47.0	66.4	96.0	100	100	100	100	93.3	63.8	49.0	24.2	71.5
PEM-LDA[38][1]	55.7	85.2	93.3	100	100	100	100	100	100	99.3	85.9	51.7	89.3
FV[39][1]	2.0	9.4	51.0	89.9	96.6	99.3	100	96.0	91.3	65.1	12.1	3.4	59.7
FV-LDA[39][1]	46.3	72.5	89.9	97.3	98.0	100	100	100	96.6	81.2	61.1	41.6	82.0
MLCE-grayscale-LDA[1]	69.1	90.6	97.3	100	100	100	100	100	100	96.6	90.6	61.1	92.1



1.a 1.b 2.a 2.b 3.a 3.b 4.a 4.b 5.a 5.b

6.a 6.b 7.a 7.b 8.a 8.b 9.a 9.b 10.a 10.b

Fig. 13. First 10 extreme pose samples in LFW database and the corresponding virtual frontal images synthesized using MLCE method. Note $*.a$ indicates the original extreme pose image and $*.b$ indicates the synthesized virtual frontal face image.

image synthesis. We categorize the facial pose into 403 discrete classes (31 $yaw\ poses \in [-75° : 5° : +75°] \times$ 13 $pitch\ poses \in [-30° : 5° : +30°]$). To conduct comprehensive evaluation, we manually labeled two subsets of LFW database, i.e. the "extreme-pose" subset and "gentle-pose" subset face images, which contains 83 and 754 images respectively. We label a face image as "extreme-pose" if the absolute yaw pose is no less than 45° or the absolute pitch pose is no less than 20°. The "gentle-pose" indicates the face images has 15°–45° absolute yaw pose or 10°–20° absolute pitch pose. Note the pose is not labeled precisely and the purpose of labeling the pose is to give more informative evaluation about the influence of pose variations for different methods.

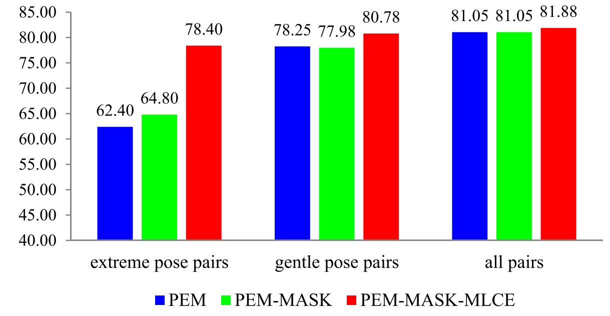


Fig. 14. MLCE evaluation on LFW database. PEM results is obtained by using original database. PEM-MASK results is obtained by using images masked by a frontal image mask. PEM-MASK-MLCE results is obtained by using automatically synthesized virtual frontal images to replace images that have large pose variations in the masked version of LFW database. Note, for each testing pair, if at least one of the image in this pair has extreme or gentle pose, the testing pair is labeled as a extreme pose or gentle pose pair.

With automatically estimated pose using five facial landmarks, if face image has gentle or extreme pose according to the aforementioned definitions, a virtual frontal face image is synthesized using MLCE method. As mentioned in Section V, there is invisible facial regions that cannot be directly synthesized and the invisible regions vary according to facial pose and the facial shape structure. To avoid training multiple verification models for different types of invisible regions, we simply use the completely visible half of the facial image and flip visible half of the face to replace the partially invisible half of the face. In this way, only one frontal facial region mask is sufficient for masking virtual frontal images. The first ten samples of face images, which have extreme pose variations in LFW and the corresponding virtual frontal face images synthesized by MLCE method are shown in Fig. 13.

The mean accuracy on LFW database are presented in Fig. 14. Note if one of the image of a testing pair belongs to extreme or gentle pose set, this pair is supposed to be a extreme or gentle pose pair. As shown, while PEM method can handle gentle-pose variation to some extent, the mean accuracy of extreme-pose pairs decline significantly. By using MLCE based virtual image synthesis, the capabilities in coping with both extreme and gentle pose data are considerably enhanced. Note as there are other complex variations exist in LFW database, when the pose problem is ideally solved the performance of extreme-pose and gentle-pose data should be similar to the average performance of all data. Since there are not many images have large pose variation in LFW database, the overall improvements on whole database is limited.

VII. Conclusion

In this paper, we aim at explicitly finding semantic correspondence between faces under different poses for face recognition across pose. To achieve this goal, we first build a statistical model named Morphable Displacement Field (MDF) from a set of 3D face models to represent rational semantic correspondence. Then, to adapt the ideal correspondence learning paradigm to face recognition application, we propose two modified objectives, i.e. image matching and probabilistic matching, for correspondence learning in face recognition scenario. As the proposed MLCE can elaborately use 3D face shape information to learn personalized semantic correspondence in 2D image space without difficult 3D reconstruction, very encouraging results of face recognition across pose are achieved. With only simple LDA classifier and intensity feature, our methods achieve or outperform state-of-the-art methods in all three typical multi-pose databases. Moreover, owe to the rational MDF regularization and probabilistic matching objective, the proposed MLCE can reliably learn the correspondence in uncontrolled "wild" environment. In future, we will work on learning the correspondences and discriminant features jointly.

References

[1] P. J. Grother, G. W. Quinn, and P. J. Phillips, "Report on the evaluation of 2D still-image face recognition algorithms," NIST, NIST Interagency/Internal, Gaithersburg, MD, USA, Rep. 7709, Mar. 2010, pp. 1–56.

[2] R. Brunelli and T. Poggio, "Face recognition: Features versus templates," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 15, no. 10, pp. 1042–1052, Oct. 1993.

[3] P. N. Belhumeur, J. P. Hespanha, and D. Kriegman, "Eigenfaces vs. fisherfaces: Recognition using class specific linear projection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 711–720, Jul. 1997.

[4] C. Liu and H. Wechsler, "Gabor feature based classification using the enhanced fisher linear discriminant model for face recognition," *IEEE Trans. Image Process.*, vol. 11, no. 4, pp. 467–476, Apr. 2002.

[5] T. Ahonen, A. Hadid, and M. Pietikainen, "Face description with local binary patterns: Application to face recognition," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 2037–2041, Dec. 2006.

[6] J. Wright, A. Y. Yang, A. Ganesh, S. S. Sastry, and Y. Ma, "Robust face recognition via sparse representation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 2, pp. 210–227, Feb. 2009.

[7] V. Blanz and T. Vetter, "Face recognition based on fitting a 3D morphable model," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 25, no. 9, pp. 1063–1074, Sep. 2003.

[8] R. Gross, I. Matthews, and S. Baker, "Generic vs. person specific active appearance models," *Image Vis. Comput.*, vol. 23, no. 11, pp. 1080–1093, Nov. 2005.

[9] X. Chai, S. Shan, X. Chen, and W. Gao, "Locally linear regression for pose-invariant face recognition," *IEEE Trans. Image Process.*, vol. 16, no. 7, pp. 1716–1725, Jul. 2007.

[10] A. B. Ashraf, S. Lucey, and T. Chen, "Learning patch correspondences for improved viewpoint invariant face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2008, pp. 1–8.

[11] S. R. Arashloo and J. Kittler, "Energy normalization for pose-invariant face recognition based on MRF model image matching," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 6, pp. 1274–1280, Jun. 2011.

[12] C. D. Castillo and D. W. Jacobs, "Wide-baseline stereo for face recognition with large pose variation," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit.*, Jun. 2011, pp. 537–544.

[13] X. Liu and T. Chen, "Pose-robust face recognition using geometry assisted probabilistic modeling," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, vol. 1. Jun. 2005, pp. 502–509.

[14] S. Li, X. Liu, X. Chai, H. Zhang, S. Lao, and S. Shan, "Morphable displacement field based image matching for face recognition across pose," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 102–115.

[15] X. Zhang and Y. Gao, "Face recognition across pose: A review," *Pattern Recognit.*, vol. 42, no. 11, pp. 2876–2896, 2009.

[16] S. J. D. Prince, J. Warrell, J. H. Elder, and F. M. Felisberti, "Tied factor analysis for face recognition across large pose differences," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 30, no. 6, pp. 970–984, Jun. 2008.

[17] A. Sharma, M. A. Haj, J. Choi, L. S. Davis, and D. W. Jacobs, "Robust pose invariant face recognition using coupled latent space discriminant analysis," *Comput. Vis. Image Understand.*, vol. 116, no. 11, pp. 1095–1110, Nov. 2012.

[18] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen, "Multi-view discriminant analysis," in *Proc. 12th Eur. Conf. Comput. Vis.*, 2012, pp. 808–821.

[19] A. Pentland, B. Moghaddam, and T. Starner, "View-based and modular eigenspaces for face recognition," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 1994, pp. 84–91.

[20] A. Li, S. Shan, and W. Gao, "Coupled bias–variance tradeoff for cross-pose face recognition," *IEEE Trans. Image Process.*, vol. 21, no. 1, pp. 305–315, Jan. 2012.

[21] Q. Yin, X. Tang, and J. Sun, "An associate-predict model for face recognition," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2011, pp. 497–504.

[22] R. Gross, I. Matthews, and S. Baker, "Eigen light-fields and face recognition across pose," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 1–7.

[23] Y. Gao, M. K. H. Leung, W. Wang, and S. C. Hui, "Fast face identification under varying pose from a single 2D model view," *IEE Proc. Vis. Image Signal Process.*, vol. 148, no. 4, pp. 248–253, 2001.

[24] A. Li, S. Shan, X. Chen, and W. Gao, "Maximizing intra-individual correlations for face recognition across pose differences," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2009, pp. 605–611.

[25] D. Gonzalez-Jimenez and J. L. Alba-Castro, "Toward pose-invariant 2D face recognition through point distribution models and facial symmetry," *IEEE Trans. Inf. Forensics Security*, vol. 2, no. 3, pp. 413–429, Sep. 2007.

[26] D. Jiang, Y. Hu, S. Yan, L. Zhang, H. Zhang, and W. Gao, "Efficient 3D reconstruction for face recognition," *Pattern Recognit.*, vol. 38, no. 6, pp. 787–798, 2005.

[27] T. Vetter and T. Poggio, "Linear object classes and image synthesis from a single example image," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 19, no. 7, pp. 733–742, Jul. 1997.

[28] A. Asthana, T. K. Marks, M. J. Jones, K. H. Tieu, and M. Rohith, "Fully automatic pose-invariant face recognition via 3D pose normalization," in *Proc. IEEE Int. Conf. Comput. Vis. (ICCV)*, Nov. 2011, pp. 937–944.

[29] W. Zhang, S. Shan, W. Gao, X. Chen, and H. Zhang, "Local Gabor binary pattern histogram sequence (LGBPHS): A novel non-statistical model for face representation and recognition," in *Proc. 10th IEEE Int. Conf. Comput. Vis. (ICCV)*, Oct. 2005, pp. 786–791.

[30] M. E. Tipping and C. M. Bishop, "Probabilistic principal component analysis," *J. Roy. Statist. Soc., Ser. B*, vol. 61, no. 3, pp. 611–622, 1999.

[31] B. D. Lucas and T. Kanade, "An iterative image registration technique with an application to stereo vision," in *Proc. IJCAI*, Aug. 1981, pp. 674–679.

[32] J. R. Shewchuk, "An introduction to the conjugate gradient method without the agonizing pain," Dept. Comput. Sci., Carnegie Mellon Univ., Pittsburgh, PA, USA, Tech. Rep. CMU-CS-94-125, 1994.

[33] T. Sim, S. Baker, and M. Bsat, "The CMU pose, illumination, and expression (PIE) database," in *Proc. 5th IEEE Int. Conf. Autom. Face Gesture Recognit.*, May 2002, pp. 46–51.

[34] P. J. Phillips, H. Moon, S. A. Rizvi, and P. J. Rauss, "The FERET evaluation methodology for face-recognition algorithms," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 10, pp. 1090–1104, Oct. 2000.

[35] R. Gross, I. Matthews, J. Cohn, T. Kanade, and S. Baker, "Multi-pie," *Image Vis. Comput.*, vol. 28, no. 5, pp. 807–813, May 2010.

[36] B. Yin, Y. Sun, C. Wang, and Y. Ge, "BJUT-3D large scale 3D face database and information processing," *J. Comput. Res. Develop.*, vol. 46, no. 6, pp. 1009–1018, 2009.

[37] Y. Taigman, L. Wolf, and T. Hassner, "Multiple one-shots for utilizing class label information," in *Proc. Brit. Mach. Vis. Conf.*, 2009, pp. 1–12.

[38] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang, "Probabilistic elastic matching for pose variant face verification," in *Proc. IEEE Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2013, pp. 3499–3506.

[39] K. Simonyan, O. M. Parkhi, A. Vedaldi, and A. Zisserman, "Fisher vector faces in the wild," in *Proc. Brit. Mach. Vis. Conf.*, 2013, pp. 1–12.

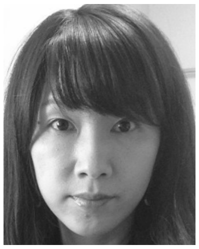
[40] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller, "Labeled faces in the wild: A database for studying face recognition in unconstrained environments," Dept. Comput. Sci., Univ. Massachusetts, Amherst, MA, USA, Tech. Rep. 7-49, 2007.

Shaoxin Li (S'14) received the B.S. degree in computer science and technology from the Huazhong University of Science and Technology, Wuhan, China, in 2009. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China.

His research interests include pattern recognition, image processing, and, in particular, face recognition and facial attribute prediction in the wild environments.

Xin Liu received the B.S. degree in software engineering from Chongqing University, Chongqing, China, in 2011. He is currently pursuing the Ph.D. degree with the Institute of Computing Technology, Chinese Academy Sciences, Beijing, China. His research interests include automatic face recognition and deep learning.

Xiujuan Chai (M'06) received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 2000, 2002, and 2007, respectively. She was a Post-Doctoral Researcher with Nokia Research Center, Beijing, China, from 2007 to 2009. She joined the Institute of Computing Technology, Chinese Academy Sciences, Beijing, in 2009, where she is currently an Assistant Professor. Her research interests include computer vision, pattern recognition, and multimodal human–computer interaction.



Haihong Zhang (M'01) received the B.E. degree from Shanghai Jiao Tong University, Shanghai, China, in 1997, and the M.E. and Ph.D. degrees from Osaka City University, Osaka, Japan, in 2001 and 2004, respectively. She is currently a Leader Research Engineer with OMRON Social Solutions Ltd., Tokyo, Japan. Her current research interests include computer vision, machine learning, and image processing.

Shihong Lao (M'05) received the B.S. degree in electrical engineering from Zhejiang University, Hangzhou, China, in 1984, and the M.S. degree in electrical engineering from Kyoto University, Kyoto, Japan, in 1988.

He has been with OMRON Corporation, Kyoto, Japan, since 1992, where he is currently an Advisory Technology Specialist with the Core Technology Center. His current interests include facial image processing, visual surveillance, and robot vision.

Shiguang Shan (M'04) received the M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and the Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS, in 2002, where he has been a Professor since 2010. He is currently the Executive Director of the Key Laboratory of Intelligent Information Processing with CAS. His research interests cover image analysis, pattern recognition, and computer vision. He especially focuses on face recognition related research topics. He has authored over 150 papers in refereed journals and proceedings in the areas of computer vision and pattern recognition, one of which received the Best Student Poster Award Runner-Up in the 2008 Computer Vision and Pattern Recognition conference. He has served as an Area Chair of many international conferences, including the 2011 International Conference on Computer Vision, the 2012 International Conference on Pattern Recognition (ICPR), the 2012 Asian Conference on Computer Vision, the 2013 IEEE International Conference on Automatic Face and Gesture Recognition, the ICPR'14, and the 2014 IEEE International Conference on Acoustics, Speech and Signal Processing. He is an Associate Editor of the IEEE TRANSACTIONS ON IMAGE PROCESSING, the *Neurocomputing*, and the *EURASIP Journal of Image and Video Processing*. He was a recipient of the China's State Scientific and Technological Progress Awards for his work on face recognition technologies in 2005.