Morphable Displacement Field Based Image Matching for Face Recognition across Pose

Shaoxin Li[1,2], Xin Liu[1,2], Xiujuan Chai[1], Haihong Zhang[3], Shihong Lao[3], and Shiguang Shan[1]

[1] Key Lab of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing, 100190, China
[2] Graduate University of Chinese Academy of Sciences, Beijing 100049, China
[3] Omron Social Solutions Co., LTD., Kyoto, Japan
{shaoxin.li,xiujuan.chai,xin.liu,shiguang.shan}@vipl.ict.ac.cn, lao@ari.ncl.omron.co.jp, angelazhang@ssb.kusatsu.omron.co.jp

Abstract. Fully automatic Face Recognition Across Pose (FRAP) is one of the most desirable techniques, however, also one of the most challenging tasks in face recognition field. Matching a pair of face images in different poses can be converted into matching their pixels corresponding to the same semantic facial point. Following this idea, given two images G and P in different poses, we propose a novel method, named Morphable Displacement Field (MDF), to match G with P's virtual view under G's pose. By formulating MDF as a convex combination of a number of template displacement fields generated from a 3D face database, our model satisfies both global conformity and local consistency. We further present an approximate but effective solution of the proposed MDF model, named implicit Morphable Displacement Field (iMDF), which synthesizes virtual view implicitly via an MDF by minimizing matching residual. This formulation not only avoids intractable optimization of the high-dimensional displacement field but also facilitates a constrained quadratic optimization. The proposed method can work well even when only 2 facial landmarks are labeled, which makes it especially suitable for fully automatic FRAP system. Extensive evaluations on FERET, PIE and Multi-PIE databases show considerable improvement over state-of-the-art FRAP algorithms in both semi-automatic and fully automatic evaluation protocols.

1 Introduction

Automatically recognizing faces in varying poses is one of the most desirable techniques in face recognition field due to its great potentials in real world applications, such as video surveillance, person re-identification and face tagging. To address the FRAP problem, a number of techniques have been proposed, which can be grouped coarsely into two categories: 3D-based and 2D-based. Please refer to the recent review paper [1] for a more detailed survey. Here we only briefly review some techniques highly related to our method.

A. Fitzgibbon et al. (Eds.): ECCV 2012, Part I, LNCS 7572, pp. 102–115, 2012.

In the category of 3D-based methods, aiming to handle pose variation by geometrical transformation, 3D face model must be estimated from 2D image at first. For instance, Blanz and Vetter [2] proposed a method to fit a full 3D morphable model to one input face image based on a statistical model of 3D human face samples. Then, obtained 3D representation was used as feature for classification. Asthana et al. [3] presented a fully automatic FRAP system. After fitting a View-based AAM [4], the input face was projected onto the aligned 3D model, which was then rotated to render a frontal view for FRAP using LGBP [5]. 3D-based methods reported impressive results on some databases. However, as 3D face recovery itself is an ill-posed problem, how to extend them to real world scenarios is still an open problem.

In the category of 2D-based methods, ill-posed 3D recovery problem was circumvented either by seeking for so-called pose-invariant features or learning to predict a virtual view in assigned pose. Gross et al. [6] proposed an eigen light-field model in which faces under different poses were represented as part of a global model containing all available pose variations. The global model could be estimated from partial model(i.e. face under some pose) and used as pose-invariant feature for face recognition. Chai et al. [7] adopted linear regression model to estimate densely sampled overlapping virtual frontal patches from corresponding non-frontal patches. Then, all virtual frontal patches were combined by averaging the overlapping pixels to form the virtual frontal face image for recognition. Prince et al. [8] exploited factor analysis model to represent faces in varying pose. Factors in different poses were tied to construct a pose-invariant "identity subspace" for final recognition.

Fundamentally speaking, the grand challenge in FRAP is an awful misalignment problem caused by the complex 3D structure of human face, i.e., the same facial point in 3D is projected to very different positions in the images of different poses in 2D. So, essentially, all FRAP methods implicitly or explicitly handle pose variation by matching pixels in 2D face images of different poses to the same semantic 3D facial points.

Recently, 2D image matching oriented FRAP methods shown great potential in handling pose variation [9–11]. These methods directly learned spatial correspondences across different poses. Ashraf et al. [9] described a data-driven approach to learn deformations between patches sampled from two different views of a face. Arashloo and Kittler [10] used MRF model to describe the matching process between two images. In their work, local patches were represented as nodes of MRF model with 2D displacement vectors as their labels. The optimal matching was found through MAP-MRF. Castillo et al. [11], employed dynamic programming-based stereo matching algorithm to find correspondences between frontal and non-frontal faces.

Although these image matching based methods achieved some success in handling pose variation, as they did not consider the correspondences between a pair of faces as a whole, correspondences obtained by these methods might not be globally conforming. Mathematically, spatial correspondences between faces in different poses can be described as a displacement field. In approach [9], as

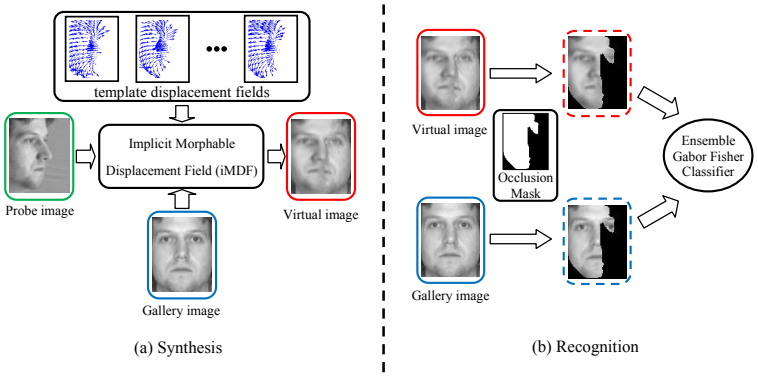


Fig. 1. Overview of the proposed method. Note that the occlusion mask is generated using the set of template displacement fields.

the deformation of each patch was learned separately, neither global conformity of displacement field over the whole face nor local consistency between adjacent patches can be guaranteed. In [10], displacements of patches were modeled by MRF model. Smooth term of MRF model did constrain neighborhood relationship but did not guarantee the global rationality of obtained displacement field. Dynamic programming-based stereo matching method introduced in Castillo et al. [11] also guaranteed only local consistency.

In this paper, we propose to build a statistical shape model to constrain the rational matching parameter which can make sure the obtained displacement field satisfying not only local consistency but also global conformity. To build the statistical model, we first generate a set of real template displacement fields from a 3D face database. Then we model target displacement field between a new pair of faces as a convex combination of these predefined template displacement fields. We name this model Morphable Displacement Field(abbr. as MDF). We further prove that the proposed MDF model also guarantees local consistency. Finally, we present an approximate but efficient solution of MDF model which not only avoids intractable optimization of the high-dimensional displacement field but also facilitates a constrained quadratic optimization. We term this solution of MDF model as implicit Morphable Displacement Field(abbr. as iMDF). Extensive evaluations on FERET [12], PIE [13] and Multi-PIE [14] databases show considerable improvement over state-of-the-art FRAP algorithms in both semi-automatic and fully-automatic evaluation protocols.

2 Method Overview

The proposed FRAP method mainly consists of two parts: virtual view synthesis via implicit Morphable Displacement Field(iMDF) and face recognition with

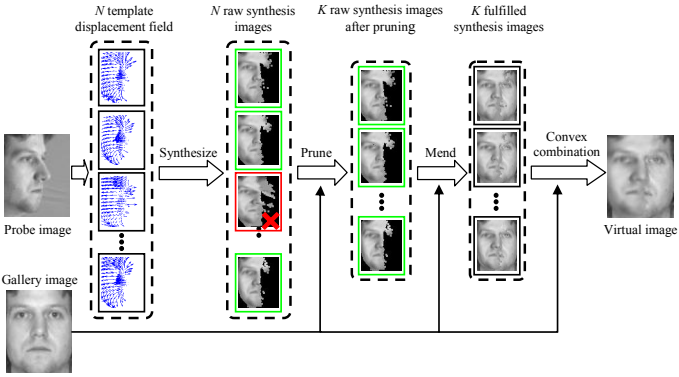


Fig. 2. Virtual view synthesis with implicit Morphable Displacement Field (iMDF)

Ensemble Gabor Fisher Classifier (EGFC) [15] (see Fig. 1). As shown in Fig. 1 (a), the virtual image is synthesized by iMDF model which can be formulated as implicit convex combination of a set of template displacement fields. Main steps of this procedure are briefly described below. Details are illustrated in Section 3 (Readers can also refer to Fig. 2).

Generate Raw Synthesis Images. With template displacement fields describing dense pixel level correspondences between face in different poses, raw synthesis image can be generated by directly taking the pixel intensity from the non-frontal probe image. We generate N template displacement fields from N 3D face models in BJUT database [16], thus N raw synthesis images can be obtained. (Detailed in section 3.1)

Prune Raw Synthesis Images. Considering computational cost, we keep only K raw synthesis images whose visible parts (non-black regions in Fig. 2) are similar to the gallery image. (Detailed in section 3.2)

Mend Undefined Parts. Raw synthesis images have some undefined parts (black regions in Fig. 2) due to self-occlusion of human face. To wipe out the influence of undefined parts, we mend these regions by directly taking corresponding pixels from the gallery image and generate K fulfilled synthesis images. (Detailed in section 3.2)

Synthesize Virtual Image. We formulate the expected virtual image as a convex combination of K fulfilled synthesis images. Optimal combination coefficients are obtained by minimizing the matching error between the expected virtual image and gallery image. (Detailed in section 3.2)

After virtual image is synthesized, pose-specific occlusion masks generated from template displacement fields are used to remove undefined regions of synthesis virtual image. And FRAP is achieved by comparing masked gallery and virtual probe image using EGFC [15], as shown in Fig. 1 (b). Detailed information is presented in Section 4.

3 Virtual View Synthesis with Implicit Morphable Displacement Field

In this section, we illustrate how we derive the virtual view synthesis method shown in Fig. 2 from the image matching point of view. The original idea is to find an optimal displacement field $T(z)$ to minimize the appearance difference between gallery image $I(z)$ and synthesis virtual image of probe image $J(z)$:

$$\arg\min_{T(z)} \sum_{z} \| I(z) - J(z + T(z)) \|_2, \tag{1}$$

where z is pixel coordinate, $J(z+T(z))$ is virtual image of probe image. Though simple and plausible, expression (1) suffers from severe overfitting due to the high dimensionality of displacement field $T(z)$. In order to obtain realistic displacement field between faces in different poses, proper restrictions must be imposed on the eligible solutions. Previous works [10] and [11] mainly focus on local consistency. In this paper, we argue that eligible solution of displacement field should possess not only local consistency but also global conformity.

3.1 Morphable Displacement Field

To obtain displacement field satisfying global conformity, we resort to template displacement fields. The template displacement field's calculation procedure is shown in Fig. 3. Note, only shape model of 3D face model is used. we draw full 3D face model including texture in Fig. 3 only for visualization purpose.

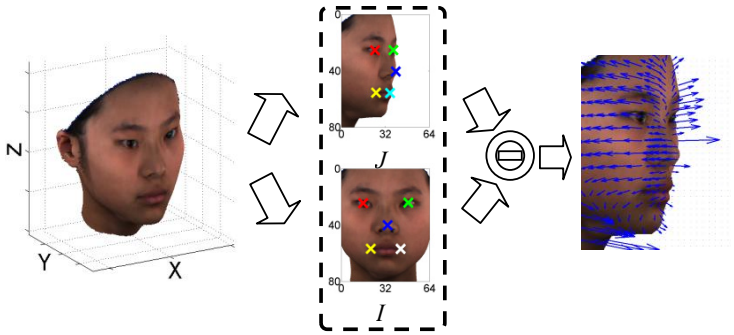


Fig. 3. Generation of template displacement field. (a) Original 3D face model; (b) Pose specific normalized 3D face model; (c) Corresponding template displacement field

Given a 3D face shape model, shown in Fig. 3 (a), one can rotate the model according to X or/and Z axis and project the rotated model onto X-Z plane to get 2D face shape models in arbitrary pose. Then with 2 or more landmarks and

assigned pose, one can further normalize the face shape models in X-Z plane using similarity transformation in order that the obtained template displacement fields and probe images are aligned. Combining 3D rotation according to X or/and Z axes and similar transformation in X-Z plane, we can generate pose specific normalized 3D face shape models, see Fig. 3 (b). Then the displacement caused by pose variation of each vertex on 3D face shape model can be calculated simply by coordinate subtraction of pose specific normalized 3D face shape models. Finally, we use a 2D discrete displacement field to approximately depict the vertices' displacement in X-Z plane caused by viewpoint variation. Since all the operators used in the calculation procedure of template displacement field T_i are linear operator, the overall process can be describe as:

$$T_i \approx L^{Frontal} \cdot S_i - L^{Pose} \cdot S_i, \tag{2}$$

where S_i is 3D face shape vector. $L^{Frontal}$ and L^{Pose} are linear operators applied to S_i in order to get pose specific normalized 3D face models shown in Fig. 3(b). Note, dense vertex-to-vertex alignment of 3D face shape model is done before we calculate template displacement fields. Thus $L^{Frontal}$ and L^{Pose} respect to different 3D face shape models are almost the same. That is to say the linear operator applied to 3D face shape model is pose-dependent but person-independent. Once we get template displacement field set $\{T_i \mid i = 1,\ 2,\ \ldots,\ N\}$, inspired by Blanz et al. [2], we propose to build a morphable model of displacement field. The morphable face model present in [2] is based on a vector space representation of faces that any convex combination of aligned shape vectors S_i of a set of exemplars describes a realistic face shape vector S:

$$S = \sum_{i=1}^{N} \alpha_i S_i; \quad s.t. \ \sum_{i=1}^{N} \alpha_i = 1,\ \alpha_i \geq 0. \tag{3}$$

According to literature [17], assuming 3D face shape vectors approximately consist a linear object class. The coefficient α_i of expression (3) will stay unchanged when linear operator L (such as 3D rigid transformation and 3D to 2D projection) is applied to shape model S:

$$S' = \sum_{i=1}^{N} \alpha_i S'_i; \quad s.t.\ S' = L \cdot S,\ S'_i = L \cdot S_i. \tag{4}$$

Combining expression (2), (3) and (4), we find that realistic displacement field between a new pair of face images can be approximately expressed as the convex combination of predefined template displacement fields:

$$\begin{aligned} T &= L^{Frontal} \cdot S - L^{Pose} \cdot S \\ &= \sum_{i=1}^{N} \alpha_i \left[L^{Frontal} \cdot S_i - L^{Pose} \cdot S_i \right] \\ &\approx \sum_{i=1}^{N} \alpha_i T_i. \end{aligned} \tag{5}$$

We term this model as Morphable Displacement Field (MDF). Modeling the target displacement field as a convex combination of template displacement fields ensures that the obtained displacement field falls in a rational parameter space. Thus the proposed MDF model is globally conforming. As mentioned before, rational displacement field should also guarantee local consistency. Local consistency indicates that spatial relationship of neighborhood pixels stays unchanged, which can be characterized as Local Order Preserving (LOP) property.

Theorem 1. *Displacement field $T(z)$ satisfies **LOP property** if for arbitrary two neighboring facial pixels (corresponding to neighboring facial points in 3D) $z_1 = (x_1, y_1)$ and $z_2 = (x_2, y_2)$, their corresponding displacement vectors $T(z_1) = (\triangle x_1, \triangle y_1)$ and $T(z_2) = (\triangle x_2, \triangle y_2)$ holds:*

$$\begin{aligned} x_1 + \triangle x_1 &\leq x_2 + \triangle x_2, if\ x_1 \leq x_2; \\ x_1 + \triangle x_1 &\geq x_2 + \triangle x_2, if\ x_1 \geq x_2; \\ y_1 + \triangle y_1 &\leq y_2 + \triangle y_2, if\ y_1 \leq y_2; \\ y_1 + \triangle y_1 &\geq y_2 + \triangle y_2, if\ y_1 \geq y_2; \end{aligned} \tag{6}$$

Here we briefly prove that in the 2D image coordinate system morphable displacement field model $T(z)$ guarantees local consistency, i.e., LOP property. Without loss of generality, we assume $x_1 \leq x_2$. For other cases shown in expression (6), the proof is similar.

Proof. $\because x_1 \leq x_2$ and all the template displacement fields $T_i(z)$ satisfy ***LOP property***, thus:

$$\begin{cases} x_1 + \triangle x_1^1 \leq x_2 + \triangle x_2^1; \\ x_1 + \triangle x_1^2 \leq x_2 + \triangle x_2^2; \\ \qquad\qquad \vdots \\ x_1 + \triangle x_1^N \leq x_2 + \triangle x_2^N; \end{cases} \tag{7}$$

where $T_i(z_1) = (\triangle x_1^i, \triangle y_1^i)$, $T_i(z_2) = (\triangle x_2^i, \triangle y_2^i)$. And combining expression (5) and (7) we derive that $T(z_1) = (\triangle x_1, \triangle y_1)$ and $T(z_2) = (\triangle x_2, \triangle y_2)$ holds: $x_1 + \triangle x_1 \leq x_2 + \triangle x_2$ □

Intuitively, template displacement fields possess LOP property naturally. And morphable displacement field formulated as a convex combination of template displacement fields also inherits LOP property. Thus the proposed morphable displacement field model fulfills the task of generating displacement field satisfying global conformity and local consistency properties at the same time.

3.2 Implicit Morphable Displacement Field

By constraining obtained Displacement Field $T(z)$ in the convex set of template displacement fields, the objective of expression (1) can be reformulated as:

$$\arg\min_{\alpha_i} \sum_z \parallel I(z) - J(z + \sum_{i=1}^{N} \alpha_i T_i(z)) \parallel_2; \quad s.t. \sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0. \tag{8}$$

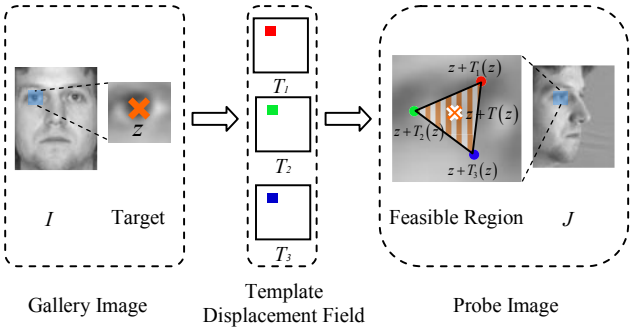


Fig. 4. Example of image matching procedure indicated by expression (8) when there are 3 template displacement fields

However, as expression (8) is not a convex optimization task, common gradient based methods [18] and [19] may fail to obtain satisfactory solution. To find a more effective way to optimize expression (8), we first investigate how it works fundamentally.

As shown in Fig. 4, each template displacement field indicates a candidate matching point $J(z+T_i(z))$ of target point $I(z)$. Since $T(z)$ is a convex combination of template displacement fields, in geometry, feasible region of all matching points is a convex hull of N vertices determined by the template displacement fields. When $N = 3$, for each target point z in image I, the feasible region of matching point is actually a triangle. Considering that, if the feasible region is sufficiently small, then grayscale of any pixels in it can be approximately interpolated by grayscale of the convex hulls vertices:

$$J(z + \sum_{i=1}^{N} \alpha_i T_i(z)) \approx \sum_{i=1}^{N} \alpha_i J(z + T_i(z)). \tag{9}$$

Actually using simple calculus it can be proved that the first order Taylor Expansion of the left and right sides of expression (9) are the same. In order to solve expression (8) more efficiently, we further relax expression (8) with (9):

$$\arg\min_{\alpha_i} \sum_{z} \parallel I(z) - \sum_{i=1}^{N} \alpha_i J(z + T_i(z)) \parallel_2; \quad s.t. \sum_{i=1}^{N} \alpha_i = 1, \alpha_i \geq 0. \tag{10}$$

The optimization problem of expression (10) is a quadratic programming. Since the objective is convex, the optimization has unique global minimum and can be solved analytically. Note that, $I(z)$ is frontal gallery image, $J(z)$ is non-frontal probe image. And parts of $J(z + T_i(z))$ are synthesized from Image J using template displacement field $T_i(z)$. As shown in Fig. 2, only visible regions in $J(z)$ can be synthesized, thus the raw synthesis images have many undefined regions. Intuitively only visible regions should be considered in the matching

procedure, in order to get rid of the influences of undefined regions we mend the black hole of raw synthesis images with $I(z)$ to generate fulfilled synthesis images, i.e., $J(z + T_i(z))$.

Compared to expression (8), rationality of expression (10) depends on the precision of approximation induced by expression (9). Therefore the size of the convex hull(see Fig. 4) should be controlled. As optimization of expression (10) is actually a common regression problem with L1 regularization adding a non-negative constraint, the optimal combination coefficients obtained by minimizing expression (10) are actually sparse. The preserved fulfilled synthesis images $J(z + T_i(z))$ with non-zero coefficients must be similar to gallery image, which indicates that the corresponding template displacement fields $T_i(z)$ are similar. Thus the corresponding convex hull shown in Fig. 4 is relatively small, which leads objective of expression (10) sufficiently close to original objective, i.e., expression (8).

Finally, for computational consideration, we prune $J(z + T_i(z))$ before the optimization and wipe out raw synthesis images which are far away from $I(z)$ in Euclidean space. Then we obtain compact version of expression (10):

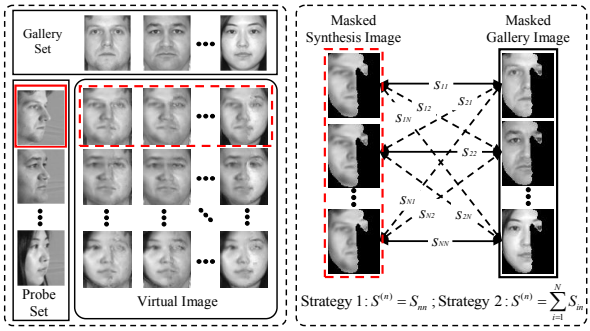
$$\arg\min_{\alpha_{N_i}} \sum_{z} \parallel I(z) - \sum_{i=1}^{K} \alpha_{N_i} J(z + T_{N_i}(z)) \parallel_2; \quad s.t. \sum_{i=1}^{K} \alpha_{N_i} = 1, \alpha_{N_i} \geq 0, \quad (11)$$

where $I(z)$ is gallery frontal image, $J(z + T_{N_i}(z))$ is the ith nearest neighbors of $I(z)$ in all raw synthesis images. In the final objective in expression (11), we implicitly obtain displacement field $T(z)$ between gallery and probe image. So we call it implicit Morphable Displacement Field (iMDF).

4 Recognition with Virtual Image

We show virtual images synthesized by proposed iMDF model using yaw $+60^\circ$ data of MultiPIE database [14] in Fig. 5(a). For each virtual image in Fig. 5(a), left facial regions are synthesized with pixels coming from probe image. We call this parts as synthesis regions. Most of right facial regions are exactly the same as gallery image, we call it mending regions. As mending regions contains no identity information, we use occlusion mask to extract only synthesis regions for subsequent recognition. The occlusion masks we use are calculated from predefined template displacement field which reflects visibility of pixels. Note, we use the same occlusion mask for different faces in the same pose, though invisible regions of different persons are slightly different.

As shown in Fig. 5(a), when probe and gallery images coming from the same person (images lie in diagonal of virtual image frame) the synthesis regions are pretty similar with true frontal face. What's more, no matter which gallery image is used as matching target, the synthesis regions of obtained virtual image always seem like the true frontal face. Fundamentally speaking, this is because displacement field generated by iMDF model satisfying both local consistency and global conformity only allows deformation caused by viewpoint variation



(a) Virtual image synthesized by iMDF (b) Recognition strategy with virtual image

Fig. 5. Image synthesis Results and Recognition Strategy

for each face. To make full use of this property, we design two kinds of recognition strategies: 1. Image-to-Image recognition; 2. Image-to-Stack recognition, as shown in Fig. 5(b). In Image-to-Image recognition strategy, similarity between gallery image I and probe image J equals to similarity between gallery image I and virtual image J', which is generated with gallery image I as matching target. In Image-to-stack recognition, we sum similarities between gallery image I and all virtual images synthesized from probe image J (for example all virtual images in red dotted frame shown in Fig. 5) as the similarity between I and J.

To make our system more robust to variations other than merely pose, we exploit EGFC introduced in [15] to calculate similarities presented in Fig. 5(b). For each pose, we train an EGFC model using masked virtual images together with corresponding masked matching targets.

5 Experimental Results

To evaluate the proposed method, we conduct FRAP experiments with a single enrolled image per person. We use frontal faces as gallery set and non-frontal faces as probe set. Totally the performance of our proposed method is evaluated in three aspects. First, we compare our method with previous FRAP algorithm to verify the effectiveness of proposed iMDF method in handling pose variation in semi-automatic evaluation protocol. Second, we present results with fully automatic configuration to show the potential of our system for automatically recognizing faces across pose. Finally, we investigate the performance bottleneck of our fully automatic system. Note, in all the following experiments, number of template displacement field $N = 700$. Preserved number of raw synthesis image $K = 25$. Image size is 64×80. In EGFC model, images are equally divided into $4 \times 2 = 8$ blocks. For each block, Gabor features with 5 scales and 8 orientations are extracted. Then PCA and LDA are employed to further extract discriminative features for final recognition. In all the experiments, dimension of PCA model is set to 300. Dimension of LDA model is set to 150 on PIE [13] and MultiPIE [14] databases and 80 on FERET [12] database.

Table 1. Semi-Automatic FRAP Performance Evaluation

(a) CMU-PIE

Method	C34	C14	C11	C29	C05	C37	C02	C22	Avg	Avg
	-60°	-45°	-30°	-15°	15°	30°	45°	60°	(C11-C37)	(C34-C22)
Chai07[7]	-	-	89	**100**	98	82	-	-	92.3	-
Castillo11[11]	56	90	**100**	**100**	**100**	**100**	96	60	**100**	87.8
Arashloo11[10]	83	91	89	91	98	**100**	95	79	94.5	90.8
GrayCCA-S1	63	99	**100**	**100**	**100**	**100**	99	84	**100**	93.1
GrayCCA-S2	63	97	**100**	**100**	**100**	**100**	99	82	**100**	92.6
GrayLDA-S1	82	**100**	**100**	**100**	**100**	**100**	99	88	**100**	96.1
GrayLDA-S2	85	**100**	**100**	**100**	**100**	**100**	99	85	**100**	96.1
EGFC-S1	78	**100**	**100**	**100**	**100**	**100**	**100**	84	**100**	95.3
EGFC-S2	**91**	**100**	**100**	**100**	**100**	**100**	**100**	**96**	**100**	**98.4**

(b) Feret

Method	bb	bc	bd	be	bf	bg	bh	bi	Avg
	-60°	-40°	-25°	-15°	15°	25°	40°	60°	
Blanz03[2]	95	95	97	**100**	97	96	95	91	95.8
Ashraf08[9]	48	70	89	96	94	82	62	42	72.9
Li09[20]	65	81	91	92	93	89	80	65	82
GrayCCA-S1	70	95	97	**100**	99	99	94	74	91
GrayCCA-S2	73	93	98	**100**	**100**	99	94	71	91
GrayLDA-S1	87	97	99	99	**100**	99	98	92	96.4
GrayLDA-S2	88	97	99	99	**100**	99	98	91	96.4
EGFC-S1	93	99	**100**	**100**	**100**	**100**	**100**	94	98.3
EGFC-S2	**98**	**100**	**100**	**100**	**100**	**100**	**100**	**98**	**99.5**

5.1 Semi-automatic FRAP Performance Evaluation

Experiments in this section are performed on two data sets, i.e., PIE and FERET. We manually label 5 landmarks (two eyes, nose tip and two mouth corners) as shown in Fig. 3. In the PIE database, we use all 68 persons with neutral expression and normal illumination at 9 yaw poses. As there are only 68 subjects, we train EGFC model in MultiPIE database. For the FERET database, We use all 200 persons at 7 different poses. The first 100 persons are used to train the EGFC model, and the rest consist the test set. Evaluation results and the pose angles we use to generate template displacement fields are given in Table 1. For comprehensive comparison we also show recognition results with features extracted by CCA and LDA from gray-scale images. Note, with identical synthesis results, "S1" uses Image-to-Image recognition strategy, while "S2" uses Image-to-Stack recognition strategy.

As shown in Table 1, our methods significantly outperform image matching based FRAP methods [9–11, 20], classical 2D-based method [7] and 3D-based method [2]. Image-to-Stack strategy is more robust than Image-to-Image strategy when use EGFC as classifier. Although we use 3D face model to calculate template displacement field, as only shape models are used, our method will not suffer from poor generalization ability [21].

5.2 Fully-Automatic FRAP Performance Evaluation

We present results of our algorithm in fully automatic evaluation protocol on the MultiPIE database. In our experiments, data from all 4 sessions with neutral expression and frontal illumination at 7 different poses are used. Images from the first 200 individuals are used as training set to train pose estimator and EGFC model, the rest 137 individuals consist test set. For each test image, two eyes are first automatically located. Then, similar to Murphy et al. [22], we extract HOG feature of the face and use SVM classifier to estimate the pose of it. The mean pose estimation accuracy of our algorithm is 95.9%. Then method introduced in [23] is used to locate eyes. Failing in detecting eye location of 2 images in totally 2100 test images, experimental results of rest data are shown in Table 2. Although EGFC algorithm is powerful in recognizing near-frontal faces,

Table 2. Fully-Automatic FRAP Performance Evaluation

Multi-PIE Fully-Auto Evaluation							
Method	08_0	13_0	14_0	05_0	04_1	19_0	Avg
	-45°	-30°	-15°	+15°	+30°	+45°	
LGBP[5]	37.7	62.5	77.0	83.0	59.2	36.1	59.3
Asthana11[3]	74.1	91.0	95.7	95.7	89.5	74.8	86.8
EGFC[15]	15.4	56.6	**99.7**	**99.3**	63.7	16.6	58.5
EGFC-S1	78.7	94.0	99.0	98.7	92.2	81.8	90.7
EGFC-S2	**84.7**	**95.0**	99.3	99.0	**92.9**	**85.2**	**92.7**

Table 3. Bottleneck Analysis of Fully-Automatic FRAP

(a) Multi-PIE Semi-Auto Evaluation							
Method	08_0	13_0	14_0	05_0	04_1	19_0	Avg
EGFC-S2	93	98.7	99.7	99.7	98.3	93.6	97.2

(b)Recognition Failed Image Distribution		Pose Estimation	
		F	T
Eye Localization	F	63	18
	T	26	1

it suffers severely from large pose variation. The mean performance of EGFC is lowest. With exactly the same experimental setting as Asthana et al. [3], the performance of proposed method is higher especially when recognition strategy 2 is used. As only 2 facial landmarks are needed, our automatic FRAP algorithm is also more potential in handling even larger pose variation automatically when faces are severely occluded (if one eye is occluded, one mouth corner and another eye can be used).

5.3 Bottleneck Analysis of Fully-Automatic FRAP

In this section, we comprehensively analyze the bottleneck of proposed fully automatic FRAP algorithm. In order to find the bottleneck, we first conduct semi-automatic evaluations on MultiPIE database [14]. Except for given ground truth eye location and face pose, all other parameters are set to be exactly the same as Section 5.2. And experimental results are shown in Table 3(a). As is shown, average recognition failure rate declines more than 50%. Totally 108 images are found to be correctly recognized by semi-automatic algorithm while falsely recognized by fully-automatic algorithm. Distribution of these recognition failed images are shown in Table 3(b). Note, if one of Normalized Root Mean Square Error(NRMSE) of two eyes exceeds 0.1, we determine that the

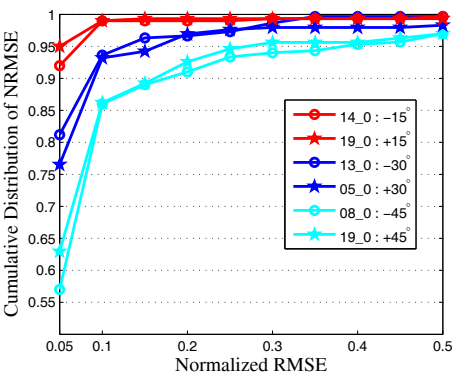


Fig. 6. NRMSE of eye localization

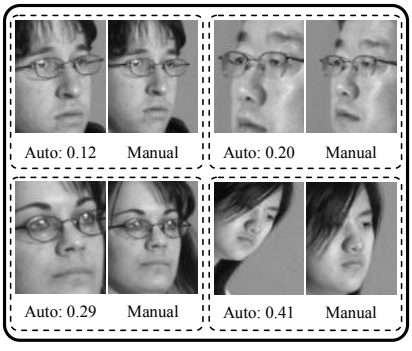


Fig. 7. Samples of false aligned images

detected eye location is false(F), otherwise true(T). Only one recognition failure occur when eye localization and pose estimation are both succeed, which suggests that our algorithm can generally handle NRMSE less than 0.1, if pose is correctly estimated. The cumulative distribution of NRMSE (we use larger NRMSE of two eyes to represent two eyes' NRMSE) of eye localization is shown in Fig. 6. Typical false auto-aligned images and corresponding manually aligned images are shown in Fig. 7 with the NRMSE value of automatic eye localization presenting under the false auto-aligned images. Since eye localization accuracy declines considerably as yaw pose angle increases, we argue that facial landmarks detection is a non-negligible bottleneck of automatic FRAP system.

6 Conclusions

This paper aims at effective image matching for pose-invariant face recognition. Given a pair of face images in different poses, we present a novel method to implicitly build the correspondences between them, which is not only locally consistent but also globally conforming. We achieve this by a convex combination of some ground truth template displacement fields generated from a 3D face database and solve it by implicit Morphable Displacement Field (iMDF). Extensive face recognition experiments on three multi-pose databases show that our method is prominent in handling large viewpoint variation. As our method relieves the burden of facial landmarks detection (2 landmarks are sufficient for our algorithm) in fully automatic FRAP system, it is potential in handling even larger pose variation when facial regions are severely occluded.

Acknowledgements. This work is partially supported by National Basic Research Program of China (973 Program) under contract 2009CB320902; Natural Science Foundation of China under contracts Nos. 61025010, 61173065, and 60832004; and Beijing Natural Science Foundation (New Technologies and Methods in Intelligent Video Surveillance for Public Security) under contract No.4111003. Haihong Zhang and Shihong Lao are partially supported by "*R&D* Program for Implementation of Anti-Crime and Anti-Terrorism Technologies for a Safe and Secure Societ", Special Coordination Fund for Promoting Science and Technology of MEXT, the Japanese Government.

References

1. Zhang, X., Gao, Y.: Face recognition across pose: A review. PR (2009)
2. Blanz, V., Vetter, T.: Face recognition based on fitting a 3d morphable model. TPAMI (2003)
3. Asthana, A., Marks, T., Jones, M., Tieu, K., Mv, R.: Fully automatic pose-invariant face recognition via 3d pose normalization. In: ICCV (2011)
4. Cootes, T., Wheeler, G., Walker, K., Taylor, C.: View-based active appearance models. In: IVC (2002)

5. Zhang, W., Shan, S., Gao, W., Chen, X., Zhang, H.: Local gabor binary pattern histogram sequence (lgbphs): A novel non-statistical model for face representation and recognition. In: ICCV (2005)
6. Gross, R., Matthews, I., Baker, S.: Eigen light-fields and face recognition across pose. FG (2002)
7. Chai, X., Shan, S., Chen, X., Gao, W.: Locally linear regression for pose-invariant face recognition. TIP (2007)
8. Prince, S., Warrell, J., Elder, J., Felisberti, F.: Tied factor analysis for face recognition across large pose differences. TPAMI (2008)
9. Ashraf, A., Lucey, S., Chen, T.: Learning patch correspondences for improved viewpoint invariant face recognition. In: CVPR (2008)
10. Arashloo, S., Kittler, J.: Energy normalization for pose-invariant face recognition based on mrf model image matching. TPAMI (2011)
11. Castillo, C., Jacobs, D.: Wide-baseline stereo for face recognition with large pose variation. In: CVPR (2011)
12. Phillips, P., Moon, H., Rizvi, S., Rauss, P.: The feret evaluation methodology for face-recognition algorithms. TPAMI (2000)
13. Sim, T., Baker, S., Bsat, M.: The cmu pose, illumination, and expression (pie) database. FG (2002)
14. Gross, R., Matthews, I., Cohn, J., Kanade, T., Baker, S.: Multi-pie. In: IVC (2010)
15. Su, Y., Shan, S., Chen, X., Gao, W.: Hierarchical ensemble of gabor fisher classifier for face recognition. FG (2006)
16. Baocai, Y., Yanfeng, S., Chengzhang, W., Yun, G.: Bjut-3d large scale 3d face database and information processing. JCRD (2009)
17. Vetter, T., Poggio, T.: Linear object classes and image synthesis from a single example image. TPAMI (1997)
18. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. IJCAI (1981)
19. Shewchuk, J.: An introduction to the conjugate gradient method without the agonizing pain. In: CMUCS-TR (1994)
20. Li, A., Shan, S., Chen, X., Gao, W.: Maximizing intra-individual correlations for face recognition across pose differences. In: CVPR (2009)
21. Gross, R., Matthews, I., Baker, S.: Generic vs. person specific active appearance models. In: IVC (2005)
22. Murphy-Chutorian, E., Doshi, A., Trivedi, M.: Head pose estimation for driver assistance systems: A robust algorithm and experimental evaluation. In: ITSC (2007)
23. Zhao, X., Chai, X., Niu, Z., Heng, C., Shan, S.: Context constrained facial landmark localization based on discontinuous haar-like feature. FG (2011)